UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Predicción de Turismo en España mediante Aprendizaje Automático Concatenado a Datos Globales de la Pandemia

Autor: Rodrigo Guerra Reyes
Tutor: David Renato Domínguez Carreta

Junio 2021

Predicción de Turismo en España mediante Aprendizaje Automático Concatenado a Datos Globales de la Pandemia

AUTOR: Rodrigo Guerra Reyes
TUTOR: David Renato Domínguez Carreta

Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2021

Resumen (castellano)

Este Trabajo Fin de Grado se ha realizado por la relevancia que ha tenido el turismo en las últimas décadas en España y el impacto que ha tenido la pandemia este último año. En este trabajo se aplican técnicas de aprendizaje automático para predecir el turismo por comunidad autónoma en España dependiendo de las defunciones causadas por el COVID-19. La falta de datos, al ser un tema muy actual, ha dificultado la búsqueda de información ya que la mayoría de las fuentes no provienen de organismos oficiales. Utilizando las herramientas de Google Colab, Python, Pandas y SkLearn se ha procedido a estudiar el turismo en España y el impacto de la pandemia.

El estudio consta de dos proyectos que pueden funcionar independientemente, pero se ha decidido unificar para obtener un resultado más complejo y útil. El primero consiste en predecir, mediante datos enfocados a la pandemia, cuántas defunciones va a haber en España los próximos meses, teniendo en cuenta la incorporación de las vacunas y los datos previos recolectados el último año. Los datos con los que se ha entrenado al algoritmo de aprendizaje automático incluyen información global de todos los países del mundo, de donde hemos hecho un filtrado ya que algunos no eran relevantes o carecían de datos. El segundo proyecto se basa en predecir el turismo en España de cada comunidad autónoma teniendo en cuenta los datos previos que tenemos desde 2016 hasta la actualidad, añadiendo en el último año los datos de la pandemia. Es claro el decremento de turismo el último año por las restricciones de movilidad. Como consecuencia de las nuevas vacunas contra el virus, el número de defunciones está bajando por lo que es posible predecir qué va a ocurrir con el turismo a lo largo del año. Se han realizado predicciones de los próximos meses las cuales pueden servir para predecir otros campos.

Palabras clave (castellano)

Aprendizaje automático, COVID-19, pandemia, turismo, vacuna, España.

Abstract (English)

This bachelor's thesis investigates the impact of the COVID-19 pandemic on the rising rate of tourism in Spain over the past decade. This investigation applies machine learning techniques to this challenge to predict tourism rates in each of the autonomous communities based on the mortality figures of Covid-19 in the given communities. The lack of data, caused by the novelty of the subject, offered challenges to the information retrieval process, and as such most of the sources used were not official statistics. Tools such as Google Colab, Python, Pandas and SkLearn were used in this study of the impact of the global pandemic on tourism in Spain.

The investigation consists of two projects that function independently from one another but were fused to achieve a more complex and useful result. The first project consists in the process of predicting (using the pandemic mortality statistics) how many deaths there will be in Spain in the coming months, considering the introduction of the various vaccines and the lack of data for the previous year. The data used to train the algorithm include global statistics, to which a filter was applied to disregard irrelevant and/or incomplete information. The second project consists in the prediction of tourism in Spain for each autonomous community, considering data collected from 2016 to the present day, adding to this the pandemic data from the previous year. This data shows a clear decrease in tourism due to the restrictions on mobility. Owing to the new vaccines against the virus, the number of deaths are declining, therefore it is possible to make predictions regarding the effect this has on tourism in the coming months, which can offer useful insights into other areas of study.

Keywords (inglés)

Machine learning, COVID-19, pandemia, tourism, vaccine, Spain.

Agradecimientos

A mi familia por haberme apoyado siempre en cualquier decisión y haber confiado en mí. A mis compañeros de la universidad por haber pasado cuatro años juntos difíciles pero divertidos gracias a ellos.

Finalmente, a todos mis amigos que han estado conmigo haciendo felices mis días y motivándome a cumplir mis metas.

LS.

INDICE DE CONTENIDOS

INDICE DE GRÁFICAS

INDICE DE TABLAS

INDICE DE ILUSTRACIONES

1 Introducción

1.1 Motivación

Esta memoria de Trabajo Fin de Grado se ha realizado por el impacto que ha tenido la pandemia en nuestra sociedad y la relevancia del turismo en España como sector crucial en la economía del país. La principal motivación es estudiar cómo ha afectado el COVID-19 al turismo. Las medidas tomadas para evitar que la pandemia no se extendiera han sido principalmente restricciones en cuanto a movilidad, afectando directamente a este ámbito relacionando el COVID-19 y el turismo.

Por estos motivos es fundamental estimar el turismo los próximos meses, aunque carezcamos de algunos datos, para poder predecir el futuro de la economía, empleabilidad y, en general, la vida en España.

Adicionalmente, una motivación personal es poder trabajar con algoritmos y herramientas de aprendizaje automático con datos reales y que pueden ser útiles para el resto. Además de proporcionar una herramienta muy versátil que se puede utilizar para cualquier otro proyecto relacionado con la pandemia.

1.2 Objetivos

El objetivo del Trabajo de Fin de Grado consiste en predecir, utilizando herramientas de aprendizaje automático, el turismo por comunidad autónoma y cómo la pandemia ha impactado en el mismo. Para conseguir resultados se ha dividido este objetivo en dos partes.

El primer objetivo a cumplir es predecir el número de defunciones causadas por el COVID-19. Este estudio es crucial para estimar el impacto que va a tener el virus los próximos meses teniendo en cuenta la introducción de las vacunas y los datos del año anterior. Este estudio aparte de ser utilizado para el turismo también se puede utilizar para cualquier otro proyecto que quiera estar relacionado con la pandemia y su impacto en otro sector. Se ha escogido el número de defunciones y no el número de casos ya que este último depende de cómo se gestiona y administra las pruebas víricas respectivas para detectar el virus. Muchos de los positivos no tienen consecuencias más allá de un malestar temporal de 2 semanas.

El segundo objetivo, el cual depende del primero, es predecir el número de turistas que va a recibir cada comunidad autónoma los próximos meses influenciados por la pandemia. Con este estudio se puede estimar la economía, empleabilidad y futuro de cada comunidad autónoma ya que muchas de estas como las islas Canarias, Baleares, Cataluña, Madrid o Andalucía dependen la mayoría de su economía del turismo.

La unión de estos dos objetivos conforma el objetivo principal que es proporcionar una herramienta capaz de predecir turismo en cada comunidad autónoma a partir de las defunciones por COVID-19.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Estado del arte**: Se realiza una explicación teórica del turismo, la pandemia y el aprendizaje automático. Los dos primeros dando un enfoque global y posteriormente centrándonos en España.
- **Diseño**: Se presenta el conjunto de datos con el que se ha trabajado y como se ha diseñado el proyecto para concatenar ambos objetivos en uno. También se explica las herramientas que se han utilizado para el proyecto.
- **Desarrollo**: Se explica el proceso llevado a cabo durante el proyecto desde la recopilación de datos inicial hasta conseguir los resultados finales. Explicando dificultades, procesos de filtrado y el uso de herramientas tecnológicas sobre aprendizaje automático.
- **Integración, pruebas y resultados**: Se muestra como se ha implementado todas las ideas previamente descritas. Las pruebas realizadas para obtener el algoritmo óptimo para este estudio y los resultados correspondientes, valorando distintos escenarios.
- **Conclusiones y trabajo de futuro**: Se opina sobre los resultados obtenidos después de haber realizado todas las pruebas y comparaciones. Se propone y motiva a continuar con el proyecto y el impacto de la pandemia en otros ámbitos.

2 Estado del arte

2.1 Turismo

2.1.1 Definición

"El turismo es un fenómeno social que consiste en el desplazamiento voluntario y temporal de individuos o grupos de personas que, fundamentalmente con motivo de recreación, descanso, cultura o salud, se trasladan de su lugar de residencia habitual a otro, en el que no ejercen ninguna actividad lucrativa ni remunerada, generando múltiples interrelaciones de importancia social, económica y cultural" (Ninoska Viloria Cedeño, "Gestión turística" pag.17) [1]. El turismo es el sector con mayor tendencia a crecer en el mundo actual, gracias a la globalización cada vez es mayor el número de viajeros, convirtiendo el turismo en uno de los pilares de muchos de estos. Las estadísticas y predicciones apuntan a que el turismo va a seguir creciendo a medida que nuevas tecnologías y nuevos métodos de transporte se vayan desarrollando, siendo cada vez más fácil y barato viajar a sitios donde hace unas décadas era inimaginable.[2] Para poder hablar de turismo debemos tener claro qué es turismo y qué no, para esto en 1937 el Consejo de la Sociedad de Naciones lo definió como "La actividad de toda persona que viaje durante veinticuatro horas o más por cualquier otro país que el de su residencia habitual". En 1993 se definió con más detalle cuales eran las características del turismo, "Las actividades de las personas que se desplazan a un lugar distinto al de su entorno habitual, por menos de un determinado tiempo y por un motivo principal distinto al de ejercer una actividad que se remunere en el lugar visitado". Por lo que cualquier persona que viaje a un lugar que no es su entorno habitual durante más de 24 horas, sin ninguna intención de remuneración, es considerado turista. Si la persona pasa menos de 24 horas se le considera excursionista. [3]

2.1.2 El turismo en España

España es el segundo país con más turistas del mundo en el 2020 solo por detrás de Francia [4]. En 2019 el turismo supuso un 12% del PIB siendo en uno de los 3 grandes sectores que impulsa la economía española. Este sector tuvo sus inicios alrededor de la década de los 60 creciendo exponencialmente hasta hoy, aunque estos últimos meses ha sido afectada por la pandemia. Al ser un pilar de la economía española es muy importante su estudio. El número de afiliados en el sector turístico durante la última década ha ido creciendo rápidamente hasta alcanzar su máximo en 2019 con 1.868.290 afiliados de promedio entre junio y agosto, el cual decremento a 1.585.914 en 2020 por culpa de la pandemia con una caída del -15,1%. [5] Teniendo en cuenta que el total de empleados en 2019 fue de 19,78 millones de personas, el turismo supuso un 9,45 % del empleo total en España. [6]

2.2 Pandemia

2.2.1 Introducción

A finales del año 2019 se empezó a alertar de un nuevo brote epidémico, una enfermedad respiratoria que nacía en Wuhan, China. Esta enfermedad se catalogó como un nuevo coronavirus al que le llamarón al principio nCoV-19, conocido también como COVID-19 o SARS-CoV-2. A inicios del año 2020 Wuhan cerró sus fronteras para contener el virus,

pero su alta contagiosidad al transmitirse por vía respiratoria mediante secreción, aerosoles o contacto directo dificultó esta medida. El 11 de marzo de 2020 la Organización Mundial de la Salud (OMS) declaró la alerta sanitaria internacional y la mayoría de los países cerraron sus fronteras para minimizar los contagios [7]. Los síntomas comunes de COVID-19 incluyen fiebre, tos y dificultades respiratorias.

2.2.2 Impacto de la pandemia

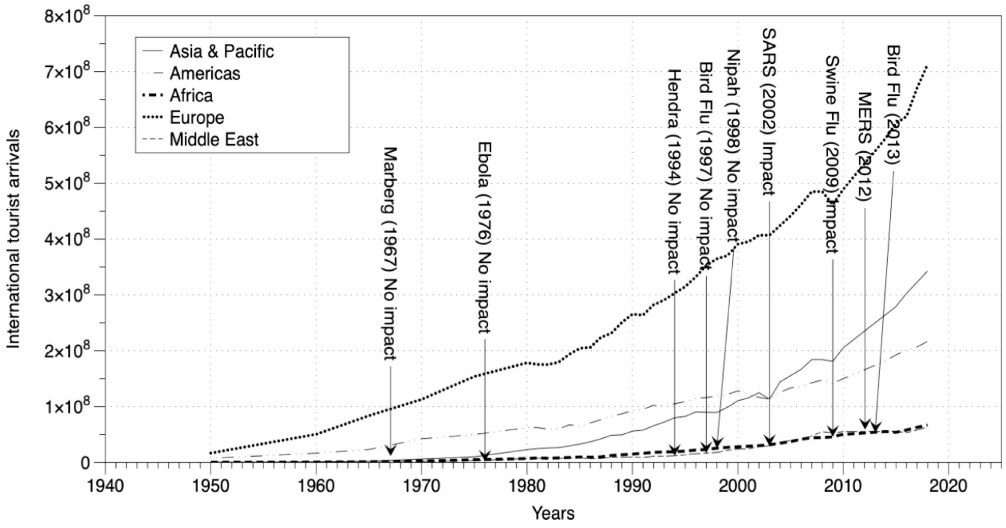
El virus ha tenido un impacto en todos los sectores socioeconómicos a nivel mundial tanto por el número de infectados como las estrictas medidas preventivas, han influido en todos los países del mundo. El COVID-19 esta entre los brotes más significativos en los últimos 53 años como podemos observar en la Tabla 2-1 (datos de abril 2020). Actualmente el número de casos COVID-19 alcanza 178.000.000 y defunciones 3.800.000.

Outbreaks	Infections	Deaths
Marberg (1967)	466	373
Ebola* (1976)	33,577	13,562
Hendra (1994)	7	4
H5N1 bird flu (1997)	861	455
Nipah (1998)	513	398
SARS (2002)	8096	774
H1N1** (2009)	762,630,000	284,500
MERS*** (2012)	2494	858
H7N9 bird flu (2013)	1568	616
COVID-19 **** (2020)	1,930,979	120,074

Tabla 2-1 Pandemias en los últimos 60 años

[8]

En los últimos 60 años hemos experimentado varios brotes epidémicos que han afectado a de diferentes maneras a la sociedad y especialmente al turismo. En la Grafica 2-1 podemos observar el número de turistas por regiones y como han afectado las principales pandemias desde 1950 hasta 2018.



Gráfica 2-1 Consecuencia de epidemias en el turismo de los continentes**.**

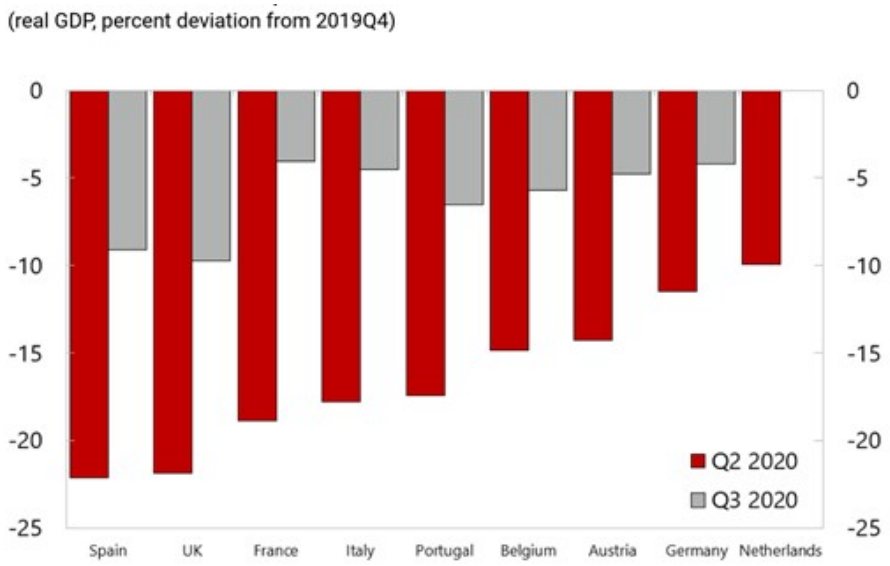
[8]

Como podemos observar en la Grafica 2-1 el turismo ha ido aumentando estos últimos años, pero tiene diferentes puntos de inflexión debido a los diferentes brotes disminuyendo el turismo en las principales regiones del planeta. La menos afectada en todos estos casos es África y una de las que más Asia, sobre todo con en el virus SARS en 2002.

El COVID-19 ha afectado a varios de los campos más importantes de la sociedad. Como todavía estamos en la epidemia no podemos obtener datos estables de la actualidad ya que van variando cada día, pero sabemos que ha tenido un impacto importante en España.

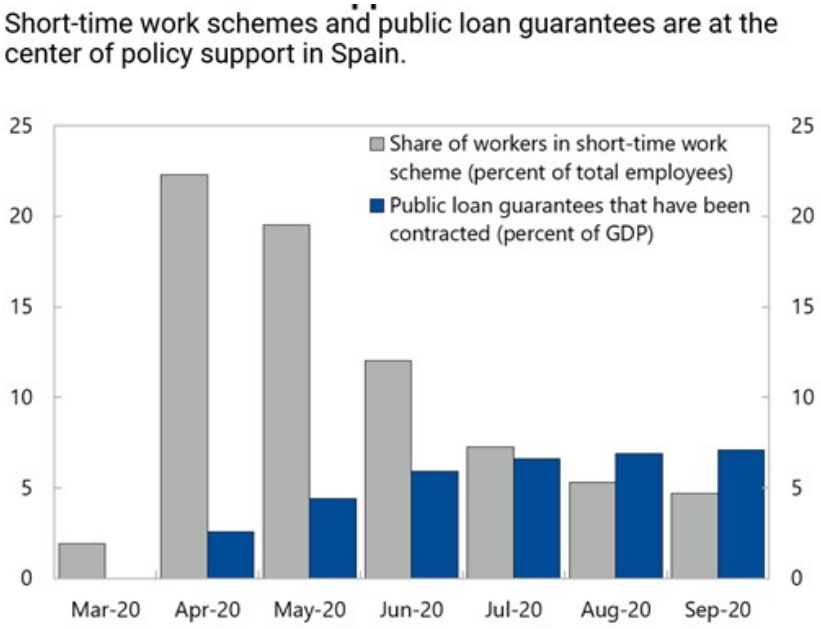
2.2.3 Impacto de la pandemia en España

En España el turismo tiene un peso de un 12% en el PIB de acuerdo con el Banco de España en 2020, es la tercera fuente de ingresos más importante del país. Cualquier cambio en este sector tiene un gran impacto en la economía como puede ser el reciente suceso del brote epidémico del COVID-19. España al ser el segundo país más visitado del mundo ha sido golpeado severamente ya que muchos de los empleos proceden del sector turístico. En el segundo cuatrimestre tuvo una bajada del PIB, por debajo del -20 comparado con 2019, siendo una de las caídas más abrumadoras de Europa. En el tercer cuatrimestre ha conseguido recuperarse gracias al turismo al abrir las fronteras.



Gráfica 2-2: Desviación del PIB en cuatrimestres 2020

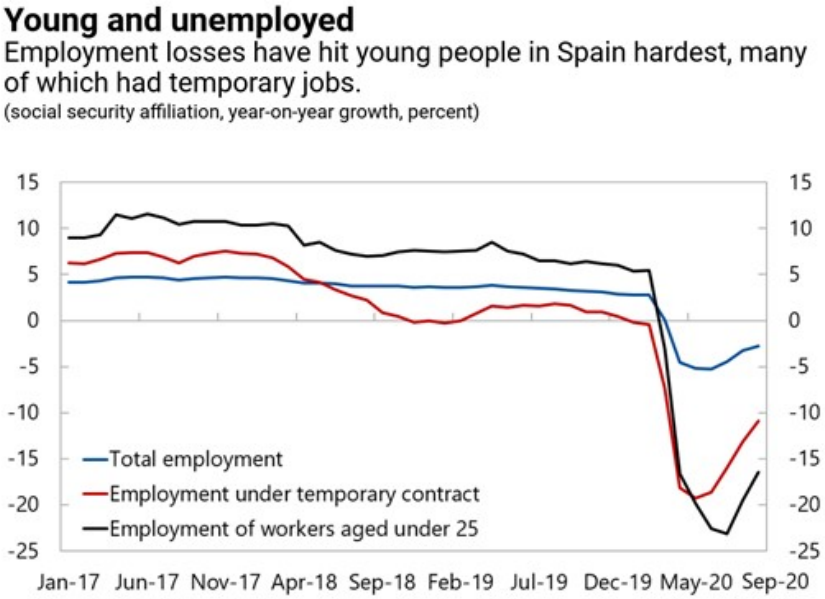
El estado español ha dado ayudas a las familias más necesitadas durante la pandemia ya que muchas de ellas carecían de ingresos durante el confinamiento. Enfocándose en los negocios pequeños, han ayudado a un 22% por ciento de la población mediante lo que conocemos como ayudas a personas en ERTE (Expediente Temporal de Regulación de Empleo).



Gráfica 2-3: Apoyo económico ante la COVID-19

La pandemia también ha afectado al sector de desempleo juvenil, uno de los peores sectores incluso antes de la pandemia, en donde, como podemos ver en la Gráfica 2-4, a partir de principios del año 2020 sufre una caída drástica tanto en los jóvenes como aquellas personas con contratos temporales, mientras que el

total de empleados decrementa, pero no tan exageradamente. Muchos de estos trabajos temporales son debido al cierre de fronteras y el bajo turismo durante esta época, por eso es crucial este estudio sobre cómo va a evolucionar el turismo en España durante los próximos meses. [9]



Gráfica 2-4 Desempleo en España 2017-2020

2.3 Aprendizaje automático

El aprendizaje autómatico deriva de la inteligencia artificial, este tipo de aprendizaje utiliza métodos algorítmicos para modelar y encontrar patrones en los datos con los que está trabajando. A raíz de estos datos el programa puede evaluar nuevas entradas con la información procesada y mejorar el rendimiento automáticamente [10]. Los datos con los que trabaja se dividen en dos tipos: los atributos de entrada y el objetivo o *target*. Los atributos de entrada son aquellos que van a ser introducidos en el modelo para poder predecir el *target* que es el resultado que queremos obtener. Existen dos tipos de aprendizajes diferenciados por el conocimiento previo que tienen sobre los datos:

- Aprendizaje no supervisado: En este tipo de aprendizaje el algoritmo realiza el procesamiento basándose únicamente en los datos de entrada, no tiene conocimiento ni del etiquetado ni de la salida esperada. El programa tiene la capacidad de aprender y relacionar los datos para crear clases propias y predecir según estas.
- Aprendizaje supervisado: Estos algoritmos tienen conocimiento tanto de los datos de entrada como de salida, conociendo de antemano la estructura de los datos. Estos algoritmos se dividen en dos fases: entrenamiento y test. Durante la fase de entrenamiento el programa puede comparar el resultado con el *target* y así calcular el porcentaje de error que está obteniendo, mejorando conforme va aprendiendo de los errores en un número de ciclos o épocas. En la fase *test* ya hemos entrenado al algoritmo e intentamos predecir los resultados con datos de entrada desconocidos aplicando el modelo entrenado. A posteriori se puede sacar conclusiones de cuanto ha sido el acierto con diferentes técnicas, una de las más comunes es la matriz de confusión en los problemas de clasificación.

Existen también dos tipos de algoritmos dependiendo de los datos de salida a obtener:

- Algoritmos de regresión: Se utiliza cuando el *target* tiene un valor continuo, por ejemplo el trabajo que estamos realizando en este proyecto tiene valores continuos ya que intenta predecir el número de turistas y defunciones. El método más utilizado para medir el rendimiento de un modelo en este tipo de algoritmos es el error cuadrático medio, Ecuación 1. También existen otros métodos como la métrica R2.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2$$

Ecuación 1 Error cuadrático medio

M: número total de patrones estimados.
i: contador de número del patrón.
real: valor real con el que comparar.
estimado: valor estimado por el modelo.

R2, es muy útil ya que podemos acotar valores entre 0 y 1. Donde 1 es el mejor resultado y 0 el peor. En algunos casos puede obtener valores negativos mostrando que el modelo se ha ajustado peor a los datos que una línea horizontal.

$$R^2 = 1 - SS_{res} / SS_{tot}$$

Ecuación 2: Ecuación Coeficiente de determinación.

SS_{res}: suma de cuadrados de los errores residuales.
SS_{tot}: suma total de todos los errores.

- Algoritmos de clasificación: Se utiliza cuando el *target* es un valor categórico o numérico pero que se pueda acotar a una clase o una etiqueta, por ejemplo, predecir si a una persona se le otorga un préstamo en base a su historial donde los resultados son "sí" o "no". El método más utilizado para medir el rendimiento de un modelo en este tipo de algoritmos es la Matriz de Confusión, Tabla 2-2. [11]

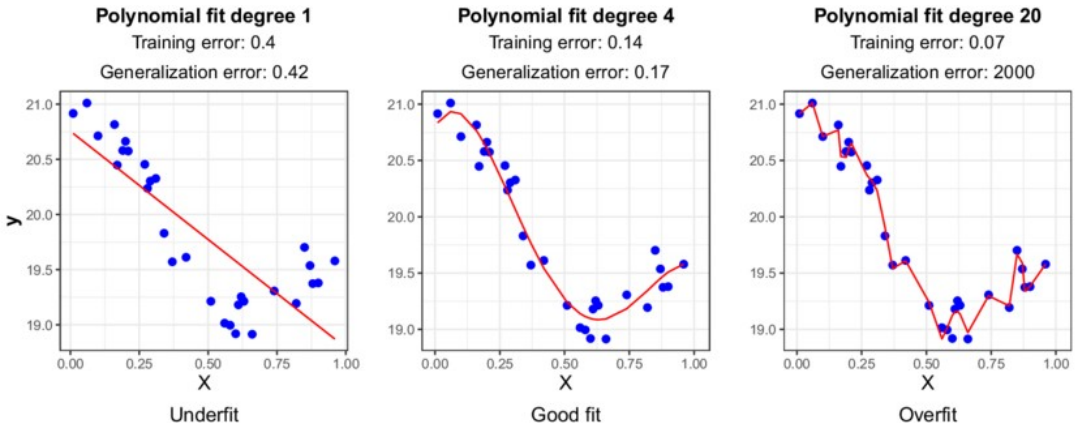
Valores Predichos\ Reales	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

Tabla 2-2 Ejemplo Matriz de confusión

VP: Verdadero Positivo; FP: Falso Positivo; FN: Falso Negativo; Verdadero Negativo.

El proceso de entrenamiento del modelo es una parte muy importante para poder obtener un resultado correcto. Un problema a la hora de entrenar el modelo es el llamado

Overfitting, el cual consiste en enseñar demasiado al algoritmo que vamos a utilizar. Como consecuencia, cuando vaya a predecir nuevas entradas que no hayan estado en la fase de entrenamiento, va a obtener un error mayor de lo esperado. También se considera a utilizar modelos más complejos de lo necesario o utilizar más atributos de los necesarios. Su contrapuesto es el *Underfitting,* el cual es más fácil de apreciar ya que en la fase de entrenamiento se obtiene un alto error a la hora de predecir concluyendo que la red no ha aprendido lo suficiente. En este proyecto nos vamos a centrar en los algoritmos supervisados y de regresión ya que contamos con valores continuos a estudiar. [12] En la Gráfica 2-5 podemos observar diferentes casos de una red entrenada con de más a menos intensidad. En la primera gráfica podemos apreciar el caso de *Underfitting* al tener una aproximación lejana al objetivo y a los datos de entrenamiento teniendo unos resultados con alto índice de error tanto en el training como en el *test*. En el centro encontramos un caso de *Good fitting* donde el modelo se ajusta bien a la tendencia de los datos, obteniendo unos resultados correctos con un menor error en el *training* pero parecido al test. La gráfica de la derecha sufre el problema de *Overfitting* al predecir perfectamente los datos de entrenamiento, pero alejarse en cuanto al error generalizado.



Gráfica 2-5: Ejemplo Underfit Good fit y Overfit **[13]**

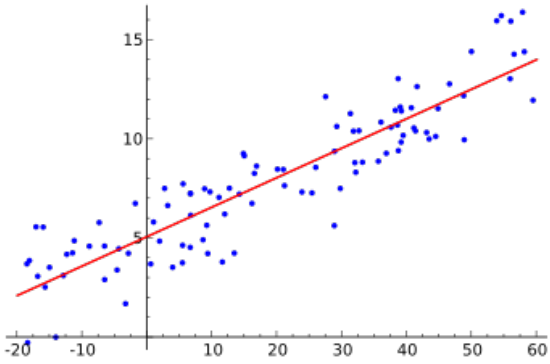
2.3.1 Regresión lineal

La regresión lineal es el algoritmo de regresión más básico y la base de la mayoría dentro del Aprendizaje Supervisado. Es utilizado para relacionar variables independientes con una variable dependiente y un término aleatorio. El modelo más básico es con solo una variable independiente en donde obtendremos una recta indicando la tendencia del conjunto de datos.

$$Y= B_0 + B_1 X_1 + B_2X_2 + \ldots + B_n X_n + E$$

Ecuación 3 Modelo de regresión lineal

Y es la variable dependiente. X_1, X_2, X_n son variables independientes. B son los parámetros del modelo también conocidos como coeficientes de regresión. B_0 es la ordenada en el origen. El algoritmo se basa en minimizar el coste de una función de error cuadrático, obtendremos unos coeficientes que nos ayudarán a obtener la recta óptima en cuanto a la regresión lineal simple, si añadimos más variables dependientes dejaría de ser una recta y podría tomar diferentes formas como una parábola. La E se refiere a las distancias entre la recta y los puntos que se pueden observar en la Gráfica 2-6. Este error corresponde al error cuadrático medio mencionado en la ecuación 1.



Gráfica 2-6 Regresión lineal básica

2.3.2 Algoritmo K-NN

El método de los k vecinos más cercanos o algoritmo K-NN es un método de clasificación que sirve para estimar la función de densidad de las entradas por la clase. Es un método no paramétrico que estima el valor de la función de densidad a posteriori de que un elemento **x** pertenezca a una clase **C** a partir de los datos de entrada, los trata como vectores que van a pertenecer a una clase. En la fase de entrenamientos el algoritmo almacena vectores y los etiqueta a la clase correspondiente. En la fase de clasificación el algoritmo intenta predecir a que clase pertenece los vectores de entrada. Mediante el cálculo de la distancia entre los vectores aprendidos en la fase anterior intenta predecir a que clase pertenece cada vector. La clase que más veces salga en la predicción entre todas las clases cercanas es la elegida para la clasificación. En la ilustración 2-1 se puede observar 5 clases y como los diferentes puntos son clasificados en cada uno por la distancia que hay entre ellos.

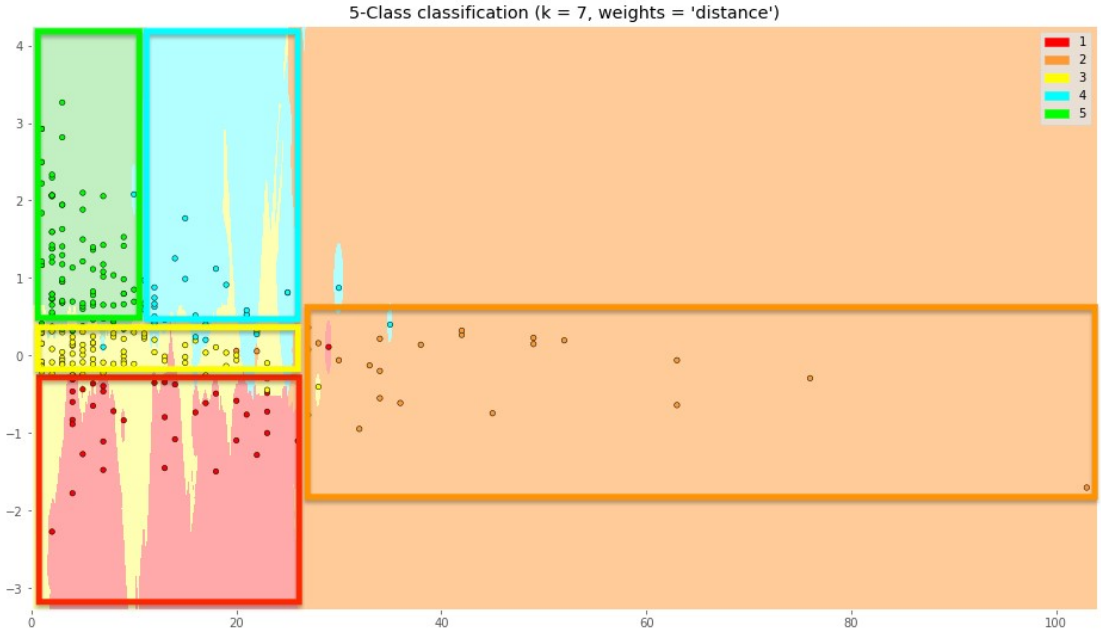
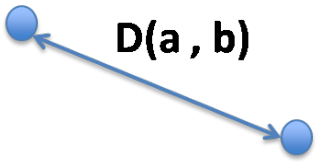


Ilustración 2-1 Ejemplo Algoritmo KNN**.**

[14]

El número **k** puede variar para obtener mejores resultados en la clasificación mediante varias técnicas como *Cross Validation*. Hay que tener cuidado con el número de **k** vecinos para intentar obtener el mejor rendimiento evitando el *overfitting* que podemos alcanzar al aumentar mucho el número de **k**. Otra de las variantes de este algoritmo es la función para calcular la distancia entre clases, la más utilizada es la distancia euclídea. Este algoritmo, aunque es famoso para problemas de clasificación también tiene su variante de regresión. [15]

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

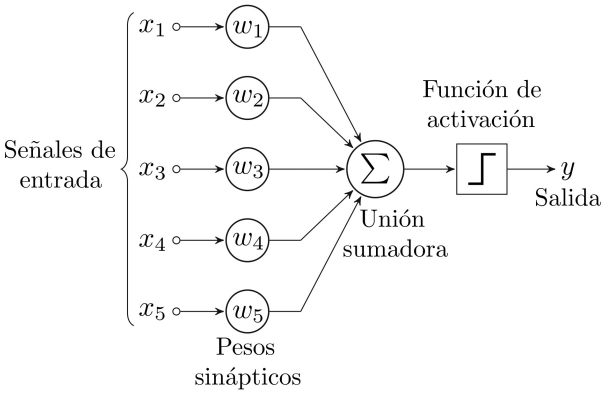


Ecuación 4 Distancia euclídea

a: es la coordenada de un punto.
b: es la coordenada de otro punto.
D: función de la distancia entre dos puntos.

2.3.3 Redes neuronales artificiales

Las redes neuronales artificiales utilizan como inspiración conceptos fundamentales de las redes neuronales biológicas como puede ser el procesamiento distribuido en unidades (neuronas) y el aprendizaje mediante modificaciones en los pesos de las conexiones. Intentan ser lo más simples posibles para facilitar su uso aun demostrando que pueden realizar tareas complejas. Todo el conocimiento se almacena en la red neuronal de una forma distribuida en los pesos de las conexiones entre las neuronas. El funcionamiento de la red empieza con la creación de una serie de elementos simples que llamamos neuronas, estas neuronas intercambian señales o información mediante conexiones. Las conexiones tienen asociado un peso por el que multiplican la señal transmitida por la neurona a otra y cada neurona aplica una función de activación al conjunto de valores de entrada que tienen para determinar su señal de salida (El sumatorio de los valores multiplicados por los pesos de todas las entradas).



X_n: dato de entrada a la neurona.
W_j: peso asociado a la conexión entre las neuronas.
y: Salida de la neurona tras hacer el sumatorio de entradas por su peso correspondiente.

Ilustración 2-2: Red neuronal simple.

Una red neuronal artificial se puede caracterizar por su patrón de conexiones, número de capas ocultas, conexiones entre las neuronas, etc. El método para determinar el valor de los pesos en las conexiones es a través de optimizadores como SGD o Adam. Otro punto importante es determinar la función de activación de las neuronas, las funciones de activación más utilizadas son la regresión lineal, sigmoide binaria y sigmoide bipolar.

Las redes neuronales pueden almacenar y recuperar datos o patrones, clasificar patrones, mapear patrones de entrada a patrones de salida, agrupar patrones por su similitud, solucionar problemas de optimización o predecir resultado basándose en comportamientos anteriores.

Las redes neuronales se pueden dividir en redes de una capa y multicapa, haciendo referencia a las capas ocultas (no incluye las de entrada o de salida) que va a tener la red.
Hay dos formas de manejar la información y actualizar los pesos de las conexiones:

- Propagación hacia delante (Feedfordward): Desde la capa de entrada hasta la capa de salida unidireccionalmente.

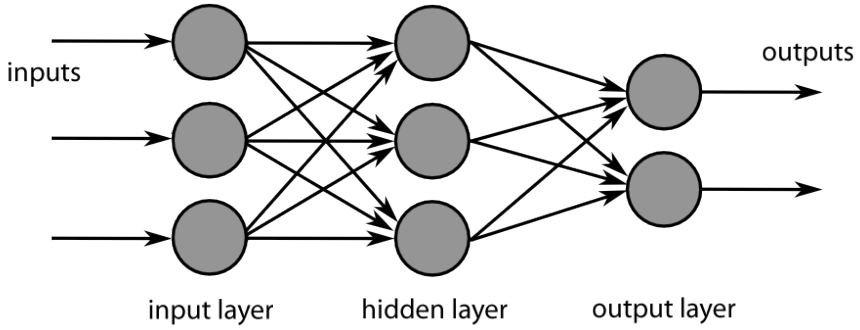


Ilustración 2-3 Red neuronal con capa oculta

[17]

- Recurrentes: Redes en las que existen conexiones de vuelta a las neuronas, es decir las neuronas reciben la señal desde la entrada y también reciben otra señal procesada por otras neuronas.

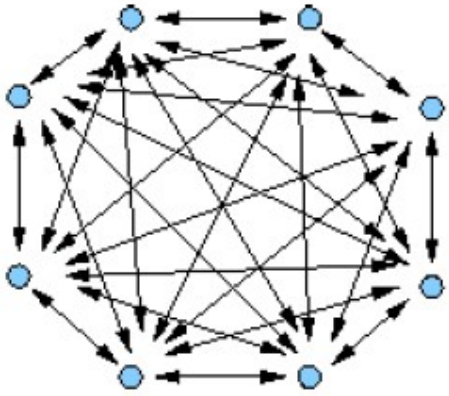


Ilustración 2-4 : Red recurrente

Las primeras redes neuronales fueron la de McCullochPitts, Perceptrón y Adaline. En este proyecto vamos a utilizar el Perceptrón multicapa debido a su gran rendimiento y optimización de recursos utilizando el backpropagation donde se utiliza un método por descenso de gradiente para minimizar el error cuadrático total de la salida. El entrenamiento consta de 3 fase:

1. Propagación hacia delante del patrón de entrenamiento de entrada.
2. El cálculo del error asociado y su retro propagación.
3. Ajuste de los pesos.

El entrenamiento puede ser costoso pero la fase de *test* solo tiene que pasar los datos de entrada por la red ya entrenada y obtener los resultados.

2.3.4 Random Forest

El algoritmo fue desarrollado por Leo Breiman y Adele Cutler, el modelo consiste en un conjunto de árboles predictores en donde cada árbol es dependiente de los valores de un vector aleatorio el cual se trata como independiente y con la misma distribución para cada árbol. Random Forest utiliza la técnica de *bagging*, por lo que cada árbol utiliza distintas partes de los datos haciendo que cada árbol se entrene con distintas muestras. Como consecuencia cuando se juntan los resultados los errores se compensan obteniendo una predicción más generalizada. [18]

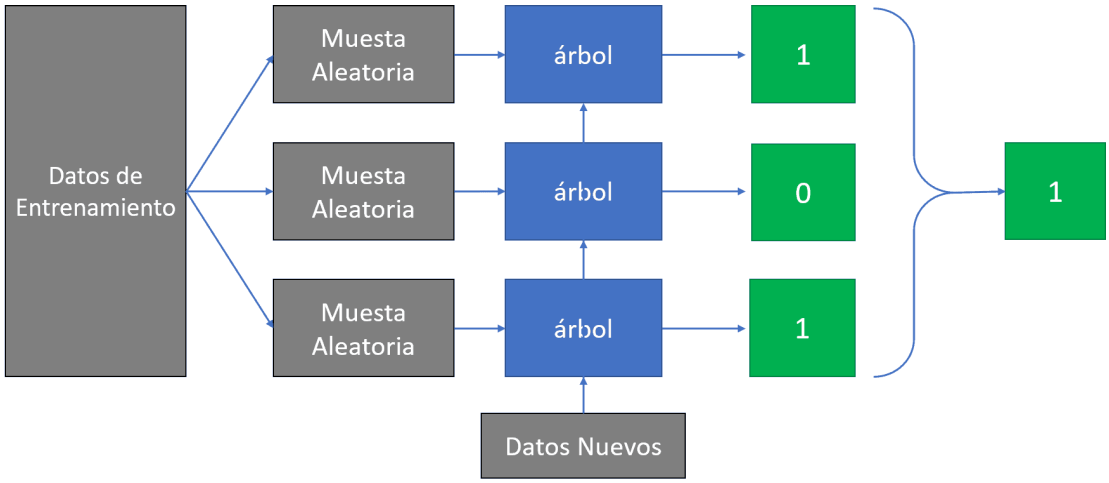


Ilustración 2-5: Ejemplo de Random Forest con 3 arboles

[19]

3 Diseño

3.1 Introducción

El problema que queremos resolver con este proyecto se compone de dos fases:

1. Predecir las defunciones causadas por el COVID-19 durante los próximos meses en España.
2. A raíz de los resultados de la Fase 1 predecir el número de turistas por comunidad autónoma en los próximos meses.

Para resolver este problema concatenado utilizamos dos modelos de aprendizaje automático con dos conjuntos de datos (*datasets)* independientes, relacionados por el factor de defunciones causadas por el COVID-19. Los datos utilizados para el problema de defunciones en España lo hemos obtenido de OWID. [20] Y los datos utilizados para predecir el turismo han sido obtenidos mayormente del Instituto Nacional de Estadísticas (INE) [21], aunque los datos se proveían separados con distintos formatos por lo que se ha tenido que juntar y normalizar para poder trabajar con ellos.

3.2 Conjunto de datos

3.2.1 Atributos y preprocesado de datos

En esta fase de preprocesado de datos se ha tratado de filtrar información irrelevante, rellenar la información incompleta por las fuentes, ya que se ha trabajado con algo tan actual ha sido muy frecuente la falta de datos en algunos periodos. Para ello se ha utilizado la función de *interpolate* de Pandas para poder obtener unos datos consistentes y continuos a través del tiempo utilizando un método lineal. Se ha normalizado los valores numéricos para poder trabajar de una forma más eficiente con el método de MinMax.

3.2.2 Defunciones COVID-19

El conjunto de datos utilizado para predecir las defunciones por COVID-19 se basan en datos relativamente estáticos de cada país como puede ser la población, PIB, índice de diabetes, pobreza. Añadiendo datos sobre la pandemia como pueden ser número de casos, vacunas, pruebas víricas, etc.… relativos a cada país. Los datos están en un intervalo de tiempo entre febrero de 2020 y hoy en día. Se accede a través de un enlace al repositorio de Github de Johns Hopkins University [22] obteniendo los datos actualizados.

Para preparar el dataset se ha filtrado los atributos para trabajar con los más relevantes para nuestra investigación y también los más consistentes ya que muchos de ellos les faltaba información o todavía no estaba actualizado para poder ser utilizado. Tras el filtrado, los atributos con los que se va a trabajar se muestran en la Tabla 3-1.

Los países con los que se ha trabajado por ser los más relevantes durante la pandemia han sido:
Australia, Brasil, España, Reino Unido, Estados Unidos, India, China, Israel, Chile, Finlandia, Francia, Alemania, Hungría, Japón, Sudáfrica, Italia, Argentina, Polonia, Colombia e Irán.

Código	Descripción
IC	Código ISO del país
C	continente
DATE	fecha
NC	nuevos casos por millón
ND	nuevas defunciones
TCPM	total de casos por millón
RR	ratio de reproducción
IPPM	pacientes UCI por millón
HPPM	pacientes en hospital por millón
NTPT	nuevos test por miles
TPC	tests por caso
TVPH	total de vacunas por cientos
PVPH	personas vacunadas por cientos
PFVPH	personas completamente vacunadas por cientos
NVSPM	nuevas vacunaciones por millón
SI	índice de restricción
P	población
PD	densidad de población
MA	edad media
A65O	por encima de 65 años
A70O	por encima de 70 años
GDP	PIB per cápita
EP	índice de pobreza
CDR	índice de muerte por problemas cardiovasculares
DP	índice de diabetes
FS	mujeres fumadoras
MS	hombres fumadores
HBPT	camas de hospital por miles
LE	esperanza de vida
HDI	índice de desarrollo humano

Tabla 3-1 Datos de COVID-19 JHU

Para más información sobre los datos OWID ha proporcionado un codebook con una explicación más detallada de los datos. (covid-19-data/owid-covid-codebook.csv at master · owid/covid-19-data (github.com))

3.2.3 Turismo España

Los datos de población [23], número de turistas [24], defunciones [25] y estudios previos sobre las defunciones en España desde 1900, se han extraído del INE. Los datos relacionados con la vacunación han sido extraídos del Centro Nacional de Epidemiología (CNE) [26] . Finalmente, los datos de defunciones y casos COVID-19 se han extraído de MoMo un equipo de investigación también gestionado por el CNE [27]. Muchos de estos datos solo se han utilizado para estudios estadísticos previos que se pueden encontrar en el Anexo B. Y se ha utilizado los datos reflejados en la Tabla 3-2 al ser los más consistentes y que tuvieran relación.

Código	Descripción
CCAA	código INE de comunidad autónoma
ND	Nuevas defunciones
NC	Nuevos casos de COVID 19
PCR	Número de casos por prueba PCR
AC	Número de casos por prueba anticuerpos
AG	Número de casos por prueba PCR
ELI	Número de casos por prueba Elisa
DESC	Número de casos por prueba desconocida
HOSP	Número de hospitalizados
UCI	Número de pacientes en UCI
NDCV	Número de defunciones por COVID-19
PFZ	Número de dosis Pfizer
MOD	Número de dosis Moderna
AZ	Número de dosis AstraZeneca
JA	Número de dosis Janssen
DEN	Número de dosis entregadas
DADM	Número de dosis administradas
PENT	Porcentaje sobre entregadas
1D	Número de personas con una dosis
COMP	Número de personas con pauta completa
TUR	Número de turistas
POP	Población

Tabla 3-2 Datos de Turismo y vacunación España

3.3 Concatenación

El problema que se quiere resolver es predecir el turismo en cada comunidad autónoma de España para los próximos meses. Para ello un factor determinante es la pandemia y las defunciones que afectan directamente al turismo al tener que cerrar fronteras y poner medidas restrictivas de movilidad. Por lo que también es necesario hacer una predicción de este dato para España. Para el modelo relacionado con las defunciones se ha elegido el algoritmo de aprendizaje automático *Random Forest* al ser el que mejor ha generalizado los datos y como veremos en el Apartado 5 ha sido el que mejor resultados nos ha dado en las pruebas. Para la predicción de turismo en cada comunidad autónoma se ha decidido utilizar el algoritmo KNN por los resultados obtenidos. Ambos enfocados a resolver un problema de regresión. En el caso de *Random Forest* utilizando RandomForestRegressor y en el caso de KNN utilizando KNeighborsRegressor de Sklearn.

El diseño planteado para afrontar todo el problema consiste en predecir primero las defunciones que va a haber en España, por ejemplo, en Julio de 2021, después hacer una ponderación teniendo en cuenta las defunciones que ha habido durante estos últimos meses de COVID-19 en cada CCAA. E introducirlo en el segundo modelo para predecir el número de turistas por CCAA.

3.4 Código aprendizaje automático

3.4.1 Python

Para el desarrollo del proyecto se ha decidido utilizar Python ya que es un lenguaje de programación interpretado que facilita la legibilidad del código. Es uno de los lenguajes de programación más flexibles, soporta la orientación de objetos y la programación funcional también conocido como programación multiparadigma. Adicionalmente, se puede utilizar Google Colab y las herramientas de aprendizaje automático más utilizadas como pueden ser Pandas, Tensorflow, Keras o SkLearn.

3.4.2 Google Colab

Google Colab es una herramienta de programación gratuita de Google. Permite ejecutar y programar en Python desde cualquier navegador con tener una cuenta gratuita de Google Drive. El formato de los ficheros donde programar y añadir información son denominados notebooks. Los notebooks de Colab son notebooks de Jupyter, facilitando el trabajo de adaptación.

Las ventajas de utilizar Colab son:

- No requiere de ningún tipo de configuración en la máquina en la que se esté trabajando.
- Acceso gratuito a GPUs.
- Permite compartir contenido fácilmente a través de Google Drive.

Los notebooks de Colab permiten combinar código ejecutable con texto informativo. En este mismo texto informativo se pueden añadir imágenes, gráficas, HTML o LaTeX.

3.4.3 Herramientas de aprendizaje automático

Pandas es una biblioteca, extensión de NumPy, para la manipulación y análisis de datos en el lenguaje de programación de Python. Pandas es una librería muy amplía donde se puede trabajar con los formatos de datos más comunes como CSV o JSON. El tipo de dato que se utiliza para la manipulación de datos con indexación integrada se llama DataFrame. La herramienta te da la facilidad de leer y escribir datos entre estructuras de dato en-memoria y formatos de archivo variados.

Ilustración 3-1: Logo de Pandas

Scikit-Learn o SkLearn es una biblioteca de aprendizaje automático desarrollada en Python y de software libre. Trabaja con bibliotecas numéricas como NumPy. Proporciona un amplio abanico de modelos para resolver problemas de clasificación, regresión o clustering. Además de herramientas para encontrar los hiperparámetros óptimos para nuestro modelo como puede ser GridSearchCV, un método de búsqueda costoso para encontrar los mejores parámetros mediante permutaciones entre todos los parámetros deseados. Finalmente, también proporciona herramientas de preproceso de datos para transformar y normalizar los datos con los que trabajar, un ejemplo puede ser MinMaxScaler para normalizar los datos de entrada entre un rango de 0 a 1. Se ha decidido utilizar scikit-learn por su simplicidad a la hora de crear un algoritmo de aprendizaje automático y su gran diversidad de herramientas para poder trabajar con Colab y pandas de una forma eficiente.

Ilustración 3-2: Logo de Sklearn

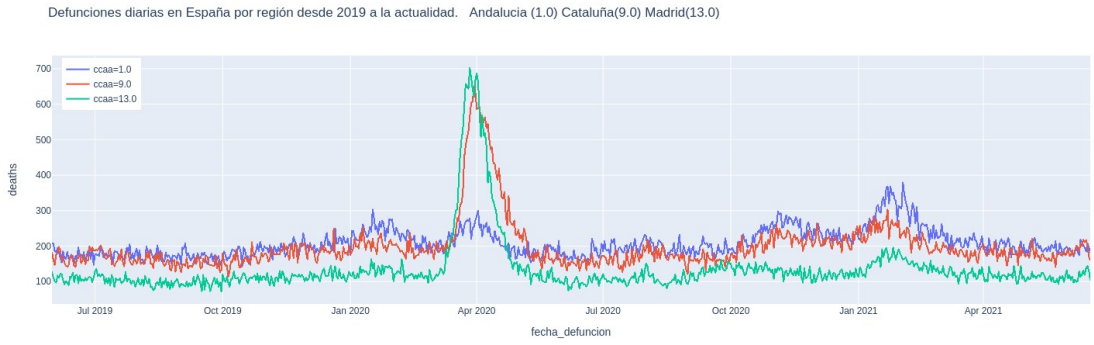
4 Desarrollo

4.1 Recopilación de datos

El proceso de recopilación de datos se ha centrado en buscar datos relacionados con el impacto de la pandemia y como ha afectado estos datos al turismo. Hemos determinado como foco de importancia las defunciones ocurridas por el virus como factor determinante a la hora de variar el número de turistas en cada región. Intenta abarcar un gran espectro de tiempo llegando a tener datos desde 1900 para comparar este virus con otras pandemias o grandes hitos en la historia. Los datos utilizados de las diferentes comunidades autónomas y del país han sido obtenidos por medio del Instituto Nacional de Estadística (INE). En cuanto a los datos sobre la pandemia y el COVID-19 se ha utilizado un estudio realizado por el Instituto de Salud Carlos III específicamente el proyecto de Sistema de Monitorización de la Mortalidad (MoMo) diaria en España. Los datos utilizados de Europa y el resto del mundo han sido obtenidos gracias a Our World In Data (OWID) que utilizaban los datos de Johns Hopkins University (JHU). Todos estos datos se han utilizado para hacer estudios previos de cómo han evolucionado las defunciones y el número de turistas en España desde hace décadas, pero solo se ha podido trabajar en los algoritmos con datos en España desde 2016 hasta marzo de 2021 y en cuanto a los datos relacionados con el COVID-19 desde febrero de 2020 hasta la actualidad.

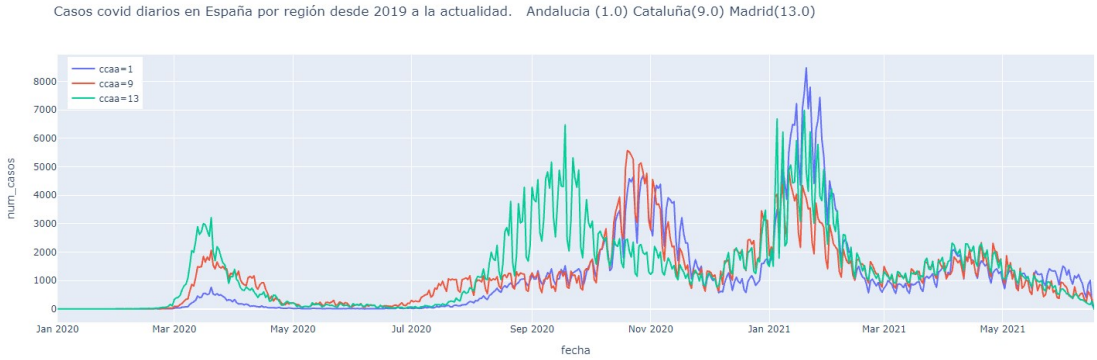
4.1.1 Visualización de los datos

La primera tarea para realizar en el estudio es ver qué factores son los más importantes y como han ido variando a través de los años/meses. En el Anexo B se pueden encontrar las gráficas respecto a defunciones según periodo, turismo, casos y defunciones COVID-19 por CCAA en España. Se han escogido las 3 comunidades más importante por población y turismo ya que mostrar las 19 haría ilegible las gráficas, Gráfica 4-1 y 4-2.



Gráfica 4-1: Defunciones diarias 2019-Actualidad. Andalucía, Cataluña y Madrid

Como conclusión del estudio previo realizado se puede observar los picos más altos de defunciones entorno a los meses de abril de 2020 y febrero de 2021. Esto se puede deber a las bajas temperaturas del invierno en donde el virus es más contagioso y las movilidades realizadas por las épocas festivas de Navidad durante los meses previos.



Gráfica 4-2: Casos COVID-19 2019-Actualidad. Andalucía, Cataluña y Madrid.

En la Gráfica 4-2 se observa un incremento de casos, llegando a tener como máximos durante el último cuatrimestre de 2020 y el primero de 2021, esto se puede deber al aumento de pruebas realizadas a individuos y a las movilizaciones tanto en verano como en Navidad de 2020 ocasionando un aumento de casos los meses posteriores. También se puede observar que cada CCAA tiene sus máximos en diferentes épocas del año coincidiendo las 3 a finales de enero de 2021.

Desde la incorporación de las vacunas contra el COVID-19 ha pasado poco tiempo para ver si están teniendo resultados en cuanto al número de defunciones o casos positivos ya que de momento está siguiendo la misma tendencia y pendiente que el año anterior de decrementar el número en el mes de mayo.

4.2 Construcción de dataset

4.2.1 Turismo en España

Para crear el dataset de España se ha tenido que juntar datos de diferentes fuentes haciendo más laborioso el proceso. Para poder juntar las tablas el primer paso ha sido establecer una asignación de atributos. Las CCAA se han normalizado con su código INE, explicadas en la Anexo C, para que en todos los datasets estén iguales ya que cada fuente las nombraba de un modo distinto. Se ha tenido que ajustar los periodos desde 2016 a la actualidad ya que ninguna fuente fiable proporcionaba datos de turismo divididos en CCAAs aparte del INE.

	ccaa	year	month	ND	NC	PCR	AC	AG	ELI	DESC
261	4	2021	4	643	3,178	1,254	0	330	0	0
190	3	2020	11	1,690	21,937	10,564	0	382	0	0
1041	16	2020	4	2,720	10,025	3,837	0	114	0	63
1116	17	2021	1	288	14,511	6,863	3	485	3	0
126	2	2021	1	1,295	24,125	18,377	0	264	1	0

Ilustración 4-1: Ejemplo Dataset turismo España

Un dato importante es que se han descartado las ciudades autonómicas de Ceuta y Melilla al obtener muy pocos datos relacionados con el turismo y la pandemia dificultando el trabajo. Por lo que no van a aparecer el código 18 ni 19 en la columna de "ccaa".

4.2.2 Defunciones por COVID-19

Este dataset ha sido más fácil de manejar ya que venía toda la información necesaria para el estudio en el mismo de JHU. Por lo que lo único se ha realizado es un filtrado de datos para solo trabajar con los relevantes para este estudio y la normalización de estos.

Los únicos datos categóricos que se encuentran en el datset son los países y los continentes por lo que se ha utilizado la técnica de *One Hot Encoding* para numerarlos y poder trabajar con ellos.

	IC	year	month	day	C	NC	NDS	RR	IPPM	HPPM	NTPT	PR	TPC
2490	FRA	2020	1	31	Europe	0.006	0.000	NaN	0.030	0.089	NaN	NaN	NaN
7321	ZAF	2021	3	30	Africa	18.725	76.714	0.94	NaN	NaN	0.432	0.043	23.0
6473	POL	2020	4	9	Europe	9.924	16.714	1.24	NaN	NaN	NaN	NaN	NaN
5881	ITA	2021	4	28	Europe	212.637	322.714	0.91	44.838	373.310	4.786	0.044	22.5
4949	ISR	2020	2	22	Asia	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
257	ARG	2020	9	14	South America	244.773	219.714	1.06	NaN	NaN	0.596	0.414	2.4

Ilustración 4-2: Ejemplo dataset Defunciones COVID-19

En la Ilustración 4-2 todavía no se ha aplicado *One Hot Encoding* para facilitar la visualización de los datos. Todos los valores NaN que aparecen son valores nulos debido a épocas donde no había empezado la vacunación o no se han obtenido todos los datos de alguna región los cuales hemos intentado tratar de ajustar a los valores previos y posteriores a esa fecha.

4.3 Aprendizaje automático

Los algoritmos que se han probado en este proyecto se han seleccionado por poder predecir problemas de regresión, además de ser los más famosos que pueden llegar a generalizar bien el problema. Para esto se han escogido los algoritmos que van a estar detallados en el Apartado 5 de pruebas. El proceso se ha basado en probar variando los diferentes hiperparametros con los que cuenta cada uno. Adicionalmente, se han utilizado herramientas auxiliares que proporciona la librería de SkLearn para poder hacer una búsqueda más detallada y/o rápida. Las dos herramientas que se han utilizado son GridSearchCV para cuando el algoritmo tenía pocos parámetros y lo utilizamos para conjuntos de datos más pequeños intentando buscar la mejor combinación posible. Y RandomSearchCV que consiste en buscar entre una serie de combinaciones pseudoaleatorias la mejor combinación, siendo mucho más rápida, y según varios ingenieros de datos casi igual de eficiente que el GridSearch. Como el proyecto se divide en dos conjuntos de datos diferentes este proceso se ha tenido que realizar dos veces para obtener el mejor algoritmo para los datos de COVID-19 en los 20 países seleccionados del

mundo y otro para los datos de turismo en España. Para el primero se ha seleccionado Random Forest y para el segundo el algoritmo KNN, las pruebas y detalles están en el Apartado 5.

4.4 Concatenación

Una vez encontrados los hiperparametros adecuados para cada uno de los algoritmos, ahora lo que nos hace falta es concatenar la información predicha por el primer algoritmo que son las defunciones por COVID-19 en España en un mes especifico, con el segundo algoritmo para predecir el turismo en ese mes en cada comunidad autónoma. Como el primer modelo no tiene los datos suficientes para especificar las defunciones por comunidad autónoma lo que se ha realizado es una ponderación teniendo en cuenta la media de defunciones que ha habido en cada CCAA en base a las del total, desde 2016 hasta 2021, para así poder hacer una ponderación específica para cada una.

ccaa	ponderación
1	0.1258920095
2	0.01701922858
3	0.02303545874
4	0.01165481188
5	0.01229636632
6	0.006233039266
7	0.1328944763
8	0.06697939293
9	0.1934625087
10	0.07018194204
11	0.01795746271
12	0.01494632116
13	0.2082655769
14	0.01666990163
15	0.005074901827
16	0.06843557415
17	0.00900102736

Tabla 4-1: Ponderaciones defunciones en España por CCAA

Adicionalmente, se ha reducido el número de casos haciendo una predicción manual con una regla de 3 para los meses en los que se ha probado este modelo teniendo en cuenta que el número de vacunación va a seguir incrementando. Los valores más estáticos como puede ser la población o el PIB no se ha modificado para las predicciones ya que su variación suponemos que será mínima los meses más cercanos. Una vez preparados todos los datos ya se puede predecir completamente las defunciones que hay en España y cómo van a afectar al turismo en cada CCAA.

5 Integración, pruebas y resultados

5.1 Comparación de modelos

Como se han utilizado dos datasets se ha tenido que realizar dos comparaciones independientes para cada uno. Ambos se han evaluado con esta serie de algoritmos de aprendizaje automático de regresión: Aleatorio, Media, Regresión Lineal (LR), Algoritmo KNN, Perceptrón Multicapa (MLP) y Random Forest (RF).

En primera instancia se ha evaluado el dataset con los datos relacionados con el COVID-19 de los 20 países internacionales. En la tabla 5-1 podemos observar los diferentes rendimientos que han dado los modelos seleccionados utilizando el error cuadrático medio como métrica. Todos con parámetros por defecto.

COVID	RMSE - Entrenamiento	RMSE – Test
Aleatorio	2,254.49	2,247.47
Media	516.59	500.39
LR	323.23	323.48
KNN	66.59	156.88
MLP	3,401.77	3,476.76
RF	20.86	23.36

Tabla 5-1: Error diferentes modelos para COVID-19.

TURISMO	RMSE - Entrenamiento	RMSE – Test
Aleatorio	1,283,205.62	1,340,569.33
Media	507,552.62	540,490.53
LR	274,618.14	306,121.90
KNN	133,030.44	163,645.61
MLP	396,263.49	383,496.56
RF	135,123.02	180,586.16

Tabla 5-2: Error diferentes modelos para Turismo en España.

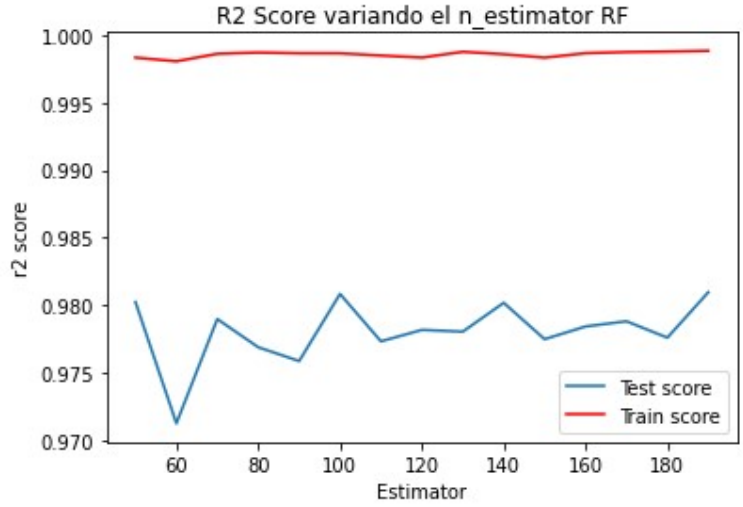
Siendo el algoritmo *Random Forest* el que mejor resultados ha dado para el primer modelo y KNN el que mejor encaja con el segundo problema.

Todas las gráficas de rendimiento, porcentaje de error de los modelos no seleccionados se encuentran en el Anexo A.

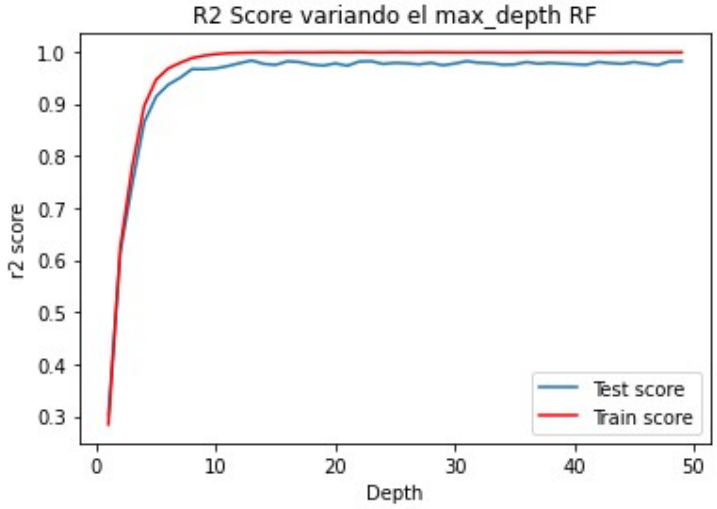
5.2 Pruebas

5.2.1 Random Forest Defunciones COVID 19

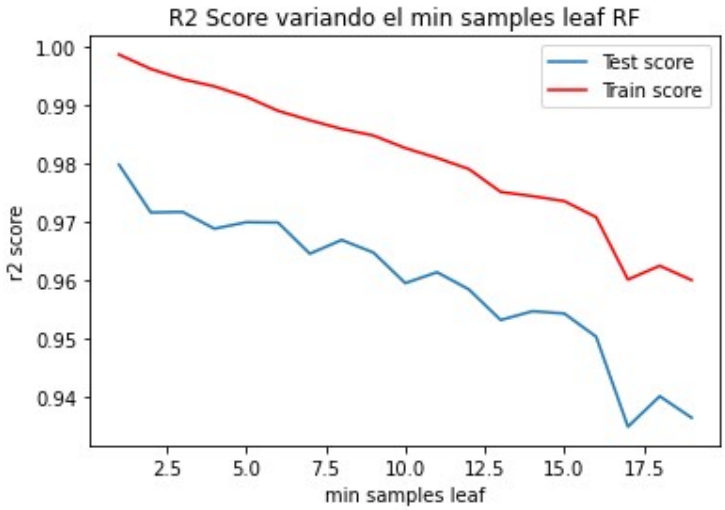
Las pruebas realizadas para el algoritmo de *Random Forest* ha sido modificar los hiperparametros más importantes como pueden ser el número de estimadores (árboles), el máximo de profundidad del árbol o el número mínimo de muestras requeridas.



Gráfica 5-1: R2 Score variando el número de estimadores RF



Gráfica 5-2: R2 score variando el máximo de profundidad RF



Gráfica 5-3: R2 score variando min_samples_leaf RF

	feat	Test score	Train score
0	auto	0.974258	0.998381
1	sqrt	0.993988	0.999227
2	log2	0.993368	0.999287

Tabla 5-3: R2_score variando max_features RF

Al obtener los resultados para cada parámetro se ha decidido escoger rangos de valores donde los resultados eran los mejores para introducir a una búsqueda de Random Search y encontrar la mejor combinación. La búsqueda de RandomSearch, no se ha hecho inicialmente ya que no se podría comparar los resultados con los obtenidos, parámetro a parámetro y observar si se ha conseguido un modelo correcto para el problema. Tras realizar el RandomSearch estos son los parámetros seleccionados:

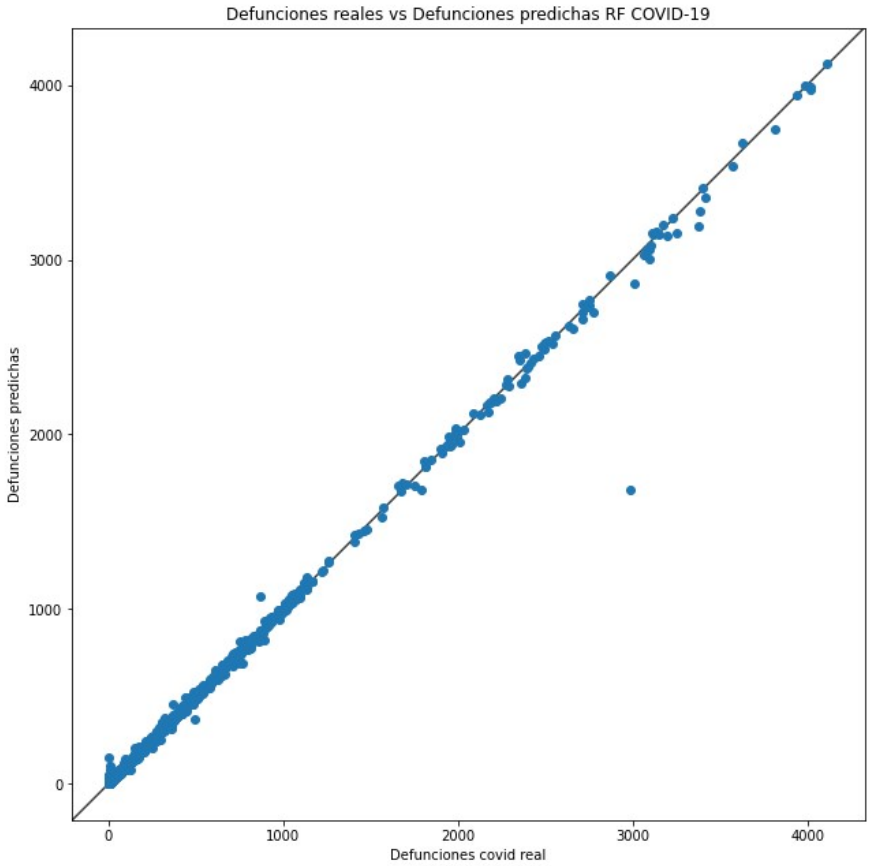
```
rf_random.best_params_
```

```
{'bootstrap': False,
 'max_depth': 90,
 'max_features': 'sqrt',
 'min_samples_leaf': 4,
 'min_samples_split': 10,
 'n_estimators': 600}
```

Ilustración 5-1: Hiperparametros para RF

Con estos hiperparametros se ha procedido a tener una serie de resultados para estimar la precisión del modelo.

Tras observar los diferentes resultados en la Gráfica 5-4 se puede apreciar que se ajusta bastante bien a la línea diagonal que es el objetivo que cumplir. Con el número de turistas reales como eje X y el predicho por el modelo en el eje Y. Se ha añadido una línea en diagonal de punta a punta para poder medir el error que obtenemos con la predicción. Poniendo un caso más práctico se dispone a predecir las defunciones en abril, julio y diciembre de 2021 en el apartado 5.3.

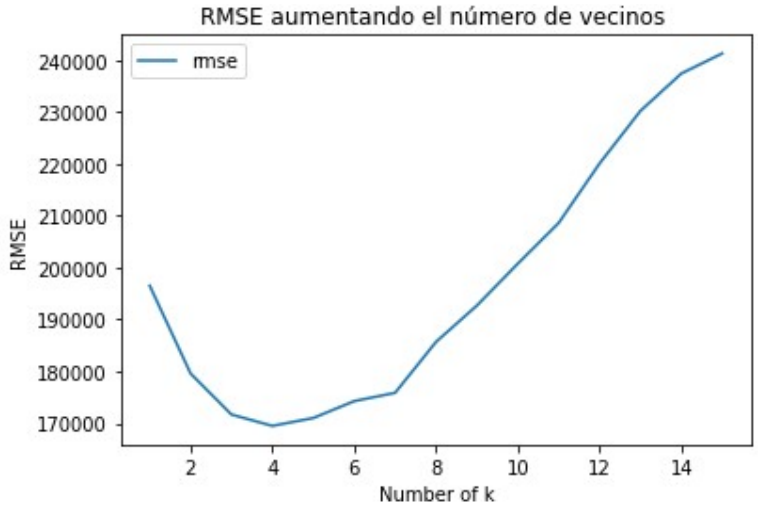


Gráfica 5-4: Defunciones predichas vs test RF

5.2.2 KNN – Turismo en España

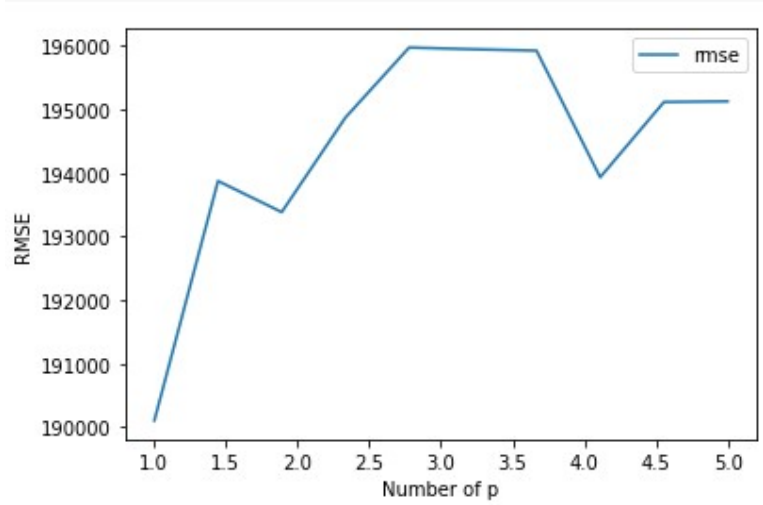
Los mejores resultados para el conjunto de datos con el objetivo de predecir el turismo en España ha sido el algoritmo KNN de regresión. Al ser uno de los algoritmos con menos hiperparametros a optimizar el estudio de este es breve y conciso.

En la Gráfica 5-5 podemos observar los resultados con la métrica del error cuadrático medio variando el número de vecinos (k) en el algoritmo. Al aumentar *k* obtenemos el mejor resultado con 4 vecinos ya que al alcanzar números altos, el algoritmo empieza a incrementar el error.



Gráfica 5-5: Variando K vs RMSE

Aumentando la variable de P de minkowski podemos observar que tiene un mejor rendimiento con valores pequeños de P.



Gráfica 5-6: Variando P de minkowski

Para poder evaluar todas las opciones posibles se ha utilizado el método de búsqueda Grid con la herramienta de SkLearn GridSearchCV para permutar por todas las opciones propuestas:

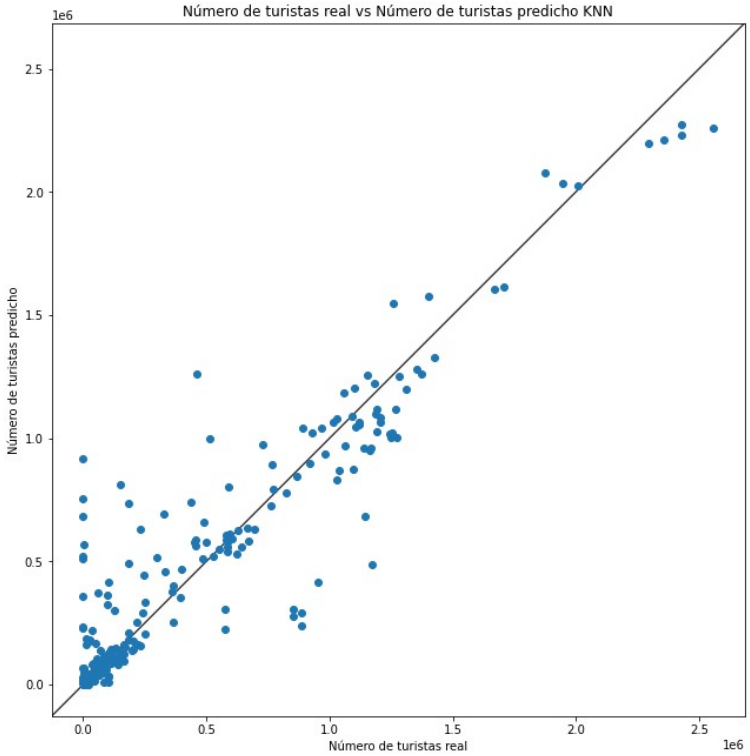
- Número de vecinos de 1 al 20
- Métricas: *manhattan*, *euclidian* y *minkowski*
- Pesos: *uniform* y *distance*.

Obtenemos estos resultados acordes al número de vecinos de 4 que hemos podido ver en la Gráfica 5-5. El mejor estimador después de aplicar una búsqueda Grid es el que podemos ver en la Ilustración 5-2, mostrando en la primera línea el resultado obtenido con la métrica R2.

```
0.9077974538973705
KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='manhattan',
                    metric_params=None, n_jobs=-1, n_neighbors=4, p=2,
                    weights='distance')
{'metric': 'manhattan', 'n_neighbors': 4, 'weights': 'distance'}
```

Ilustración 5-2: GridSearchCV KNN en Turismo España

Los resultados con esta configuración para el algoritmo KNN la podemos ver en la Gráfica 5-7. El error tiende más a estimar un número superior de turistas al esperado, podemos suponer que esto se debe a que tenemos más información de los meses con turismo constante desde 2016 a 2020 y el modelo no ha ajustado correctamente los valores de la pandemia y su efecto negativo en cuanto al número de turistas.

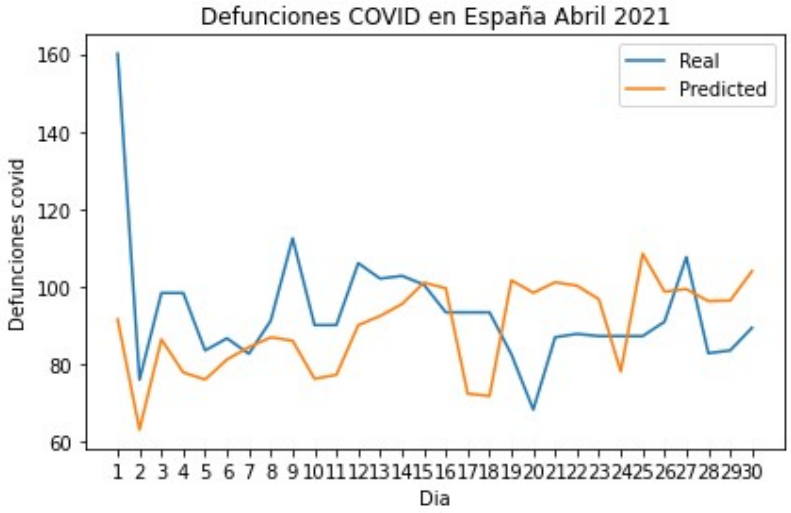


Gráfica 5-7: Predicho vs real KNN

5.3 Resultados

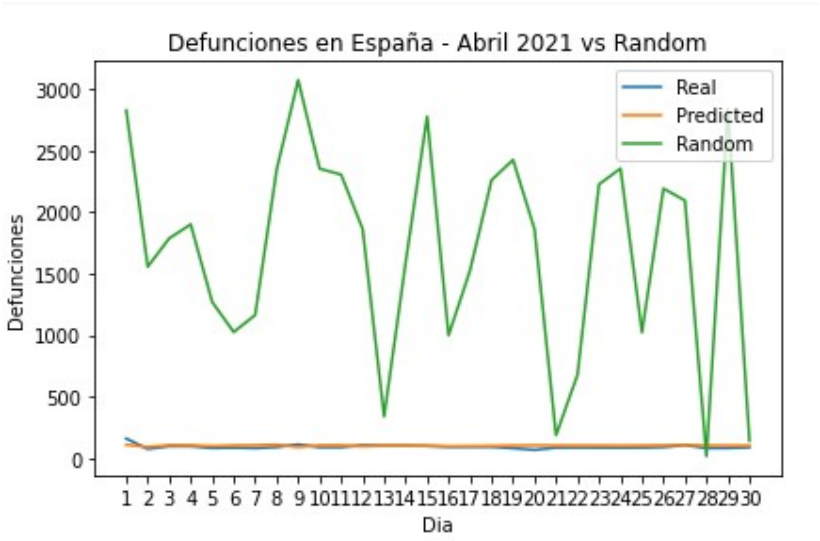
5.3.1 Predicción abril 2021

Este problema es el más cercano temporalmente, del cual tenemos datos con los que contrastar nuestra predicción. Como podemos ver en la Gráfica 5-8 el programa hace una buena estimación de cómo va a evolucionar el mes de abril, teniendo en cuenta que al principio tiene un *outlier* de 160 defunciones los datos reales, el resto del mes se ajusta bastante a la realidad. Como se observa en la Grafica 5-13 predice mejor que una serie de números aleatorios entre el mínimo y el máximo o la media de los resultados, Gráfica 5-14.

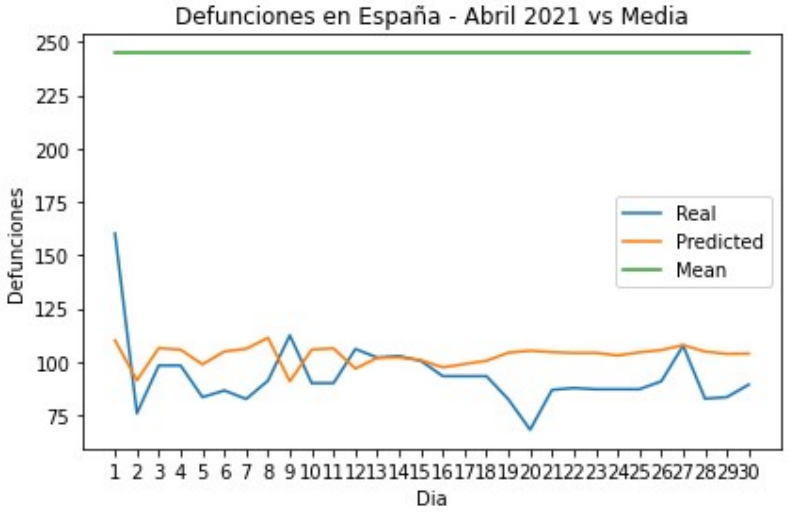


Gráfica 5-8: Defunciones predichas vs reales COVID-19 abril 2021

Gráfica 5-9: Defunciones por COVID-19 durante 2021. Comparado con lo predicho y lo real

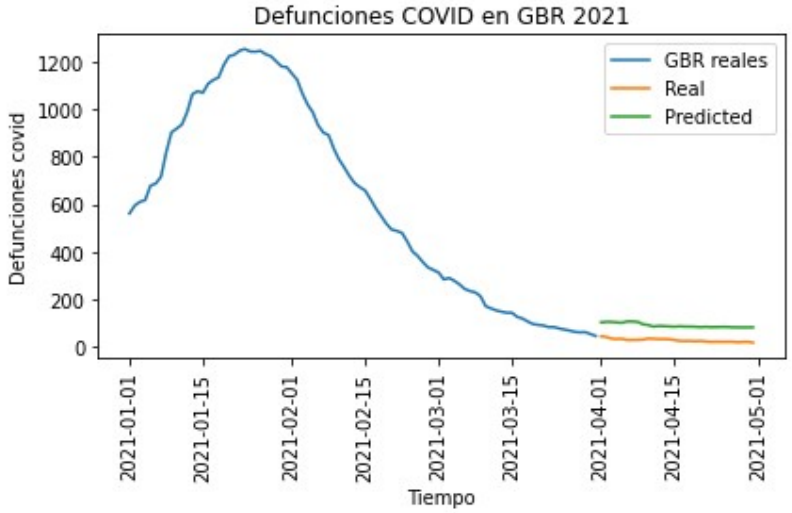


Gráfica 5-10: Defunciones predichas vs random vs reales COVID-19 RF

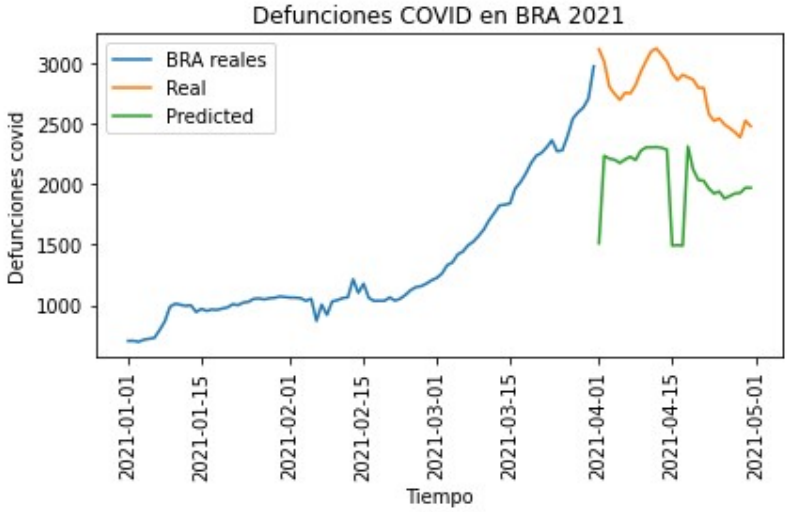


Gráfica 5-11: Defunciones predichas vs Media vs reales COVID-19 RF

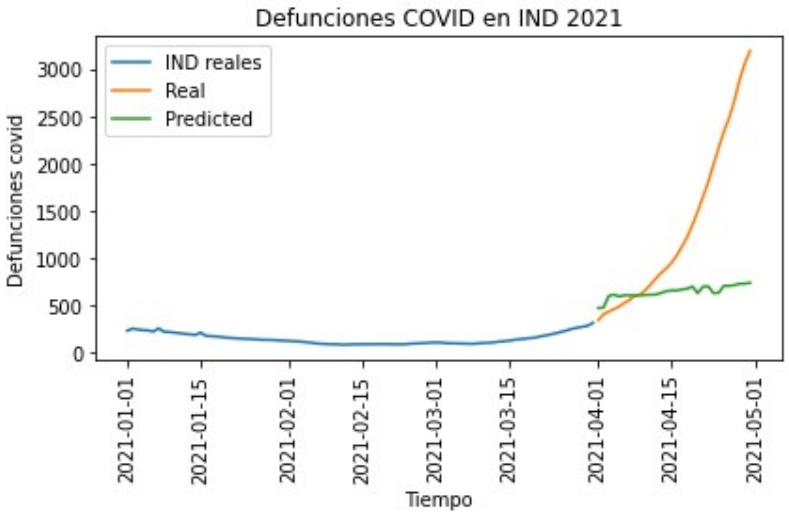
Se ha realizado también las predicciones para las defunciones por COVID-19 en diversos países influenciados por la pandemia como son Reino Unido, Brasil o India. Con esto se quiere comparar el error en diferentes escalas y ver si el modelo está prediciendo correctamente, diferenciando cada país y sus circunstancias.



Gráfica 5-12: Defunciones COVID-19 abril en Reino Unido.



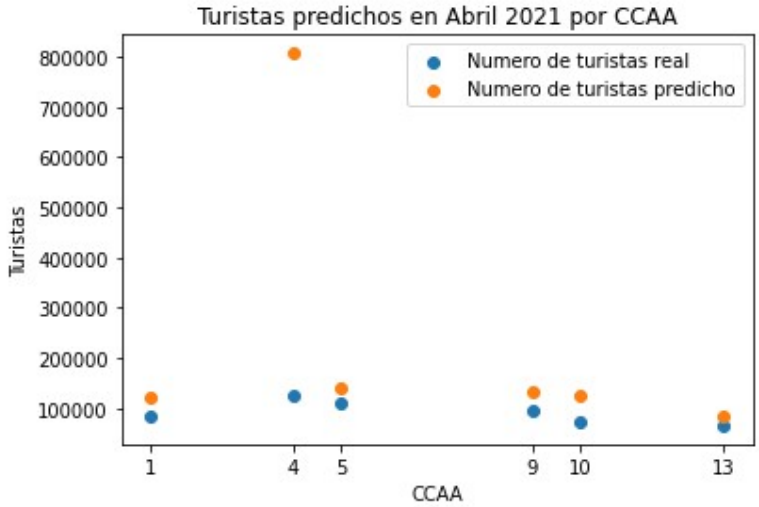
Gráfica 5-13: Defunciones COVID-19 abril Brasil



Gráfica 5-14: Defunciones COVID-19 abril India

Como podemos observar la tendencia de los meses anteriores se mantiene en el ejemplo de Reino Unido, Gráfica 5-12. En la Gráfica 5-13 con Brasil predice por debajo de la realidad teniendo en cuenta que los valores de los meses anteriores están por debajo de los valores reales de abril. Finalmente, India ha tenido un despunte el mes de abril el cual no se ha podido predecir por el modelo, que ha seguido con la tendencia de los meses anteriores, Gráfica 5-14.

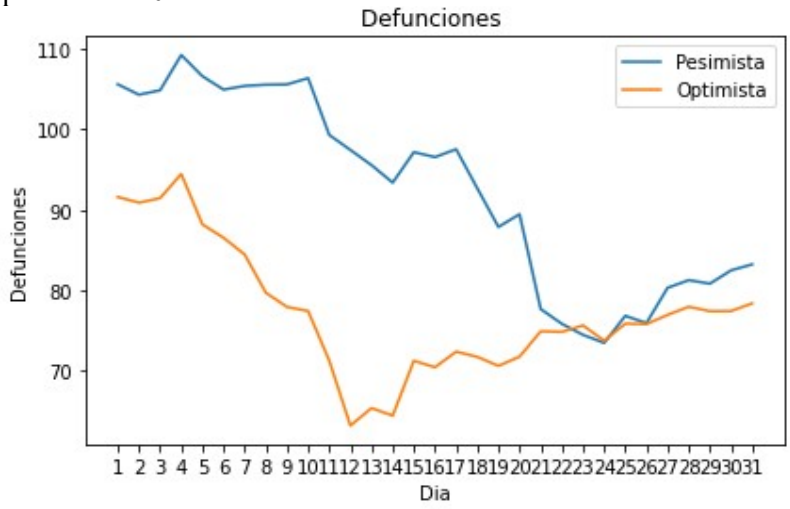
Posteriormente vamos a predecir nuestro verdadero objetivo que es el número de turistas por CCAA. Al introducir los datos predichos en el primer modelo y utilizar los datos de entrada reales de abril podemos observar en la Gráfica 5-15 que se ajusta bastante a las 5 comunidades (Andalucía, Baleares, Canarias, Cataluña, Valencia y Madrid). Teniendo en cuenta que la que más se diferencia es Baleares con una predicción de 800.000 turistas en abril, un número que sería normal si no estuviéramos en una pandemia. El resto encaja con el decremento de turismo debido al virus y tiene un menor error.



Gráfica 5-15: Turistas predichos por CCAA

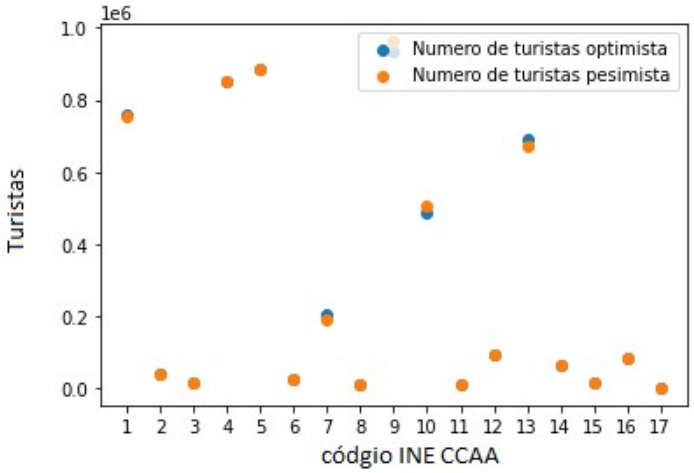
5.3.2 Predicción Julio 2021

El problema planteado ahora es Julio de 2021 un mes del cual todavía no se tiene información con la que contrastar. Se ha decidido utilizar dos escenarios: uno optimista en donde España va a alcanzar en torno a un 90% de vacunación (al menos una dosis) y otro donde se alcanzará un 60% de vacunación. Teniendo en cuenta que actualmente en junio estamos superando el 40% de vacunación.



Gráfica 5-16: Defunciones predichas para España en julio 2021

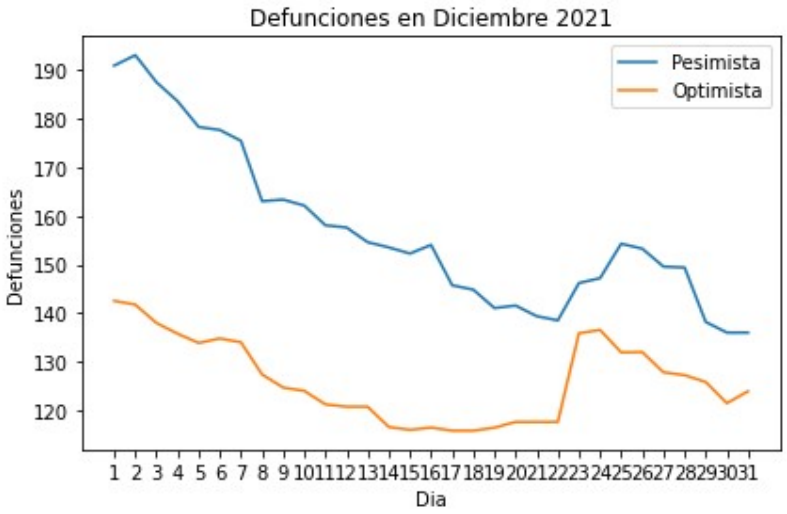
Aunque en la Gráfica 5-16 se ha podido observar una gran diferencia entre el escenario optimista y el pesimista en la Gráfica 5-17 no observamos apenas diferencia. Podemos concluir que el número de defunciones por COVID-19 no está influyendo en el modelo para predecir el turismo por CCAA de la forma en la que esperábamos. Ya sea porque en verano de 2020 se abrieron las fronteras y se redujo las restricciones de movilidad, salvando un poco el turismo. Por lo que este año la vacunación durante los meses de verano, que son los que menos casos y defunciones ha habido en general, no afecta tanto al turismo y durante los próximos meses se va a mantener como el año pasado.



Gráfica 5-17: Predicción optimista vs pesimista

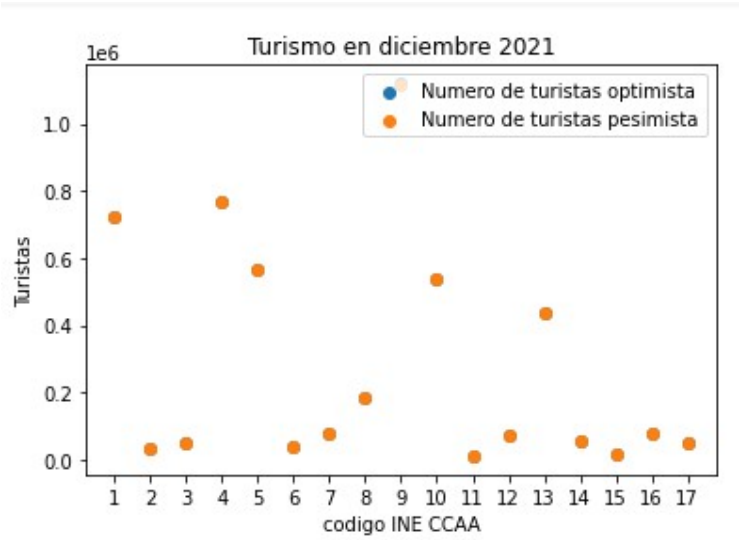
5.3.3 Predicción Diciembre 2021

El último problema planteado es diciembre de 2021, mes donde hay mucha movilidad por los festivos y la Navidad. Como en el problema anterior se observa una diferencia entre el escenario pesimista en donde llegaríamos sobre el 80% de vacunación y el optimista con un 100% de vacunación. Teniendo en cuenta que en ambos casos se ha conseguido la vacunación de rebaño.



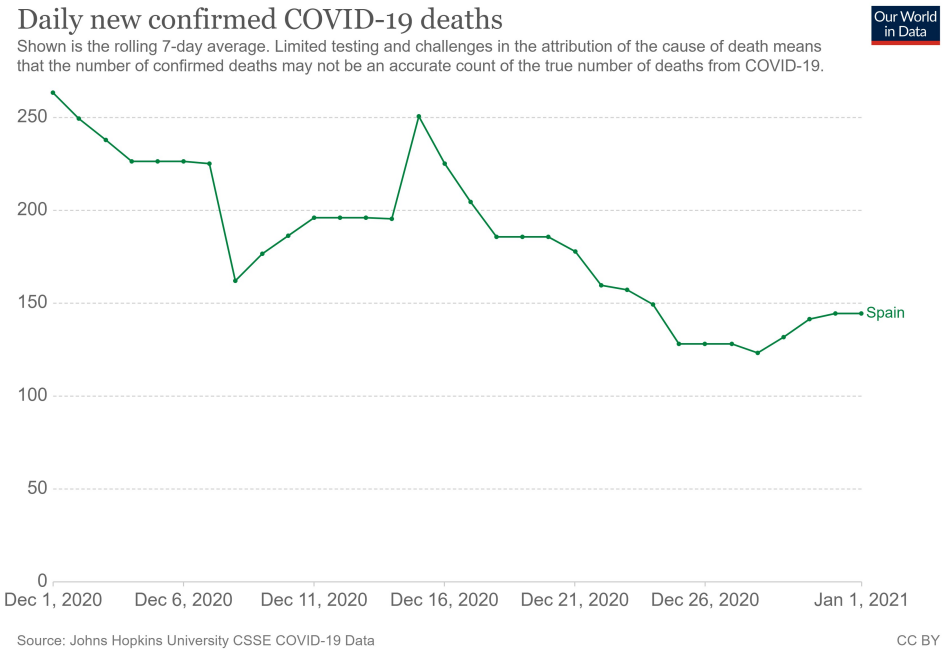
Gráfica 5-18: Defunciones COVID- 19 diciembre 2021

En la Gráfica 5-19 podemos observar que casi no va a afectar las defunciones por COVID-19 al turismo en diciembre de 2019 teniendo los mismos resultados para el escenario optimista como el pesimista.



Gráfica 5-19: Turismo España diciembre 2021

Viendo estas Gráficas y comparando con los datos del 2020 proporcionados por OWID en la Gráfica 5-18. El número de defunciones en diciembre de 2021 se va a reducir pero no muy significadamente, esto puede ser porque todavía no se ha observado el impacto de las vacunas con certeza.



Gráfica 5-20: Defunciones diciembre 2020 OWID

Y los datos recolectados por el INE en el Anexo A del cual se ha capturado los datos para comparar con diciembre de 2019 el turismo de 3 CCAA: Andalucía, Cataluña y Madrid. De donde podemos concluir que el turismo en diciembre de 2021 va a ser parecido al que hubo en 2019.

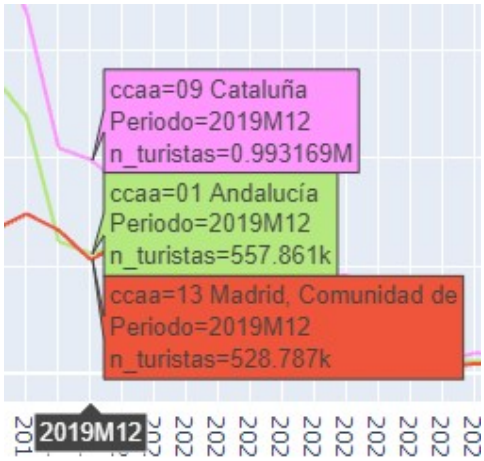


Ilustración 5-3: Turismo diciembre 2019

6 Conclusiones y trabajo futuro

6.1 Conclusiones

El Trabajo de Fin de Grado propuesto es una idea innovadora y actual la cual tiene como dificultades el tratado de datos y la concatenación de resultados acumulando el error de un algoritmo a otro. Las pruebas y resultados obtenidos han sido correctos con lo esperado, pero tienen un margen de mejora introduciendo nuevos datos que vayan publicándose a lo largo de los meses para tener un modelo mejor entrenado. Todavía es muy pronto para poder tener en cuenta el impacto de las vacunas tanto en la mortalidad del virus como en el turismo. Los algoritmos que mejor han funcionado son *Random Forest* para el problema de predecir las defunciones en España y el algoritmo KNN para predecir el turismo en cada comunidad autónoma. Obteniendo un cierto error como podemos observar en la predicción de abril 2021 ya que el modelo no puede predecir cuándo las restricciones de movilidad se levantan o se imponen. Tampoco tiene muchos datos de la vacunación ya que esta se ha empezado a utilizar estos últimos meses y no tenemos datos suficientes para predecir con detalle el impacto que va a tener hacia el turismo. Es una herramienta que le faltan datos pero puede conseguir ser muy útil para el análisis de la pandemia en cualquier sector.

6.2 Trabajo futuro

Este proyecto deja abierto un gran abanico de posibilidades como proyectos para análisis del impacto de la pandemia, ya que no solo se predice el turismo que va a haber en España, sino que también deja como herramienta la predicción de defunciones por COVID-19 en cualquier país del mundo.

En cuanto a este proyecto se puede mejorar probando con diferentes algoritmos o herramientas de aprendizaje automático como pueden ser Keras o Tensorflow que pueden llegar a un nivel de abstracción menor consiguiendo diferentes resultados. Adicionalmente, a medida que pasan los meses se está consiguiendo una información más verídica y fiable sobre la pandemia y los resultados de la vacunación por lo que sería interesante continuar con este proyecto para predecir cómo va a seguir evolucionando España o cualquier otro país hasta que se acabe esta epidemia.

Referencias

[1] V. Cedeño, N. Elena. "Desarrollo turístico y su relación con el transporte." (2012).
[2] D. Bella, M. Gurría. "Introducción al turismo." (2019).
[3] J. L. A, López, (2015). Definiciones: turismo-turista. Papers de turisme, (14-15), 17-25., página 19-20
[3] UNWTO "World Tourism Barometer and Statistical Annex, December (2020) | World Tourism Organization"
[4] T. González. (2020). La descoordinación y las soluciones regionalizadas, los errores del verano | Economía (hosteltur.com)
[5] R. Fernandez. (2020). Statista. Número de personas con trabajo en España de 2010 a 2021. https://es.statista.com/estadisticas/501094/empleados-en-espana/
[6] T. T. Wang, H. S. Moon, A., Le, & N. Panchal, (2020). Proceedings of the OMS COVID-19 Response Conference. Journal of Oral and Maxillofacial Surgery, 78(8), 1268-1274.
[7] , M. Škare, D. Riberio Soriano, M. Porada-Rochoń, Impact of COVID-19 on the travel and tourism industry, Technological Forecasting and Social Change, Volume 163, (2021),120469,ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2020.120469. (https://www.sciencedirect.com/science/article/pii/S0040162520312956)
[8] N. Arregui, L. Liu y W. Oman (14 de noviembre de 2020) "Cinco gráficos sobre la economía española y respuesta de España a la COVID-19" Departamento de Europa del FMI.
[9] D. B., Millán, J. G.,Boticario,& P. I., Viñuela (2006). Aprendizaje automático. Sanz y Torres.
[10] S. E. Regalado Bolaños y A. E. Bautista Ulcuango, «Análisis, implementación y evaluación de modelos de aprendizaje automático relacional,» Quito: UCE, 2019.
[11] D. M., Hawkins (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.
[12] Badillo, Solveig & Banfai, Balazs & Birzele, Fabian & Davydov, Iakov & Hutchinson, Lucy & Kam-Thong, Tony & Siebourg-Polster, Juliane & Steiert, Bernhard & Zhang, Jitao David. (2020). An Introduction to Machine Learning. Clinical Pharmacology & Therapeutics. 107. 10.1002/cpt.1796. Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, pp. 114-119 vol.1, doi: 10.1109/IJCNN.2000.857823.
[13] Na8, (2018). Algoritmo k-Nearest Neighbor | Aprende Machine Learning
[14] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng. (2017). Learning k for kNN Classification. ACM Trans. Intell. Syst. Technol. 8, 3, Article 43 (April 2017), 19 pages. DOI:https://doi.org/10.1145/2990508 para knn
[15] A. Cartas - Trabajo propio, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=41534843
[16] H. Hellknowz (2019) - MultiLayerNeuralNetwork_english.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=11397827.
[17] T. Hastie, R. Tibshirani, & J. Friedman (2009). Random forests. In The elements of statistical learning (pp. 587-604). Springer, New York, NY.
[18] J.M. Heras. (2020) Random Forest: combinando árboles.
[19] E. Dong, H. Du, L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

[20] Instituto Nacional de estadística (INE, 2020)
[21] E Dong, H Du, L Gardner. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1
[22] INE. (2021). Número de turistas según comunidad autónoma de destino principal.
[23] INE. (2021). Población por comunidades y ciudades autónomas y tamaño de los municipios.
[24] INE. (2021). Movimiento Natural de la población: Defunciones. Por lugar de residencia (Serie desde 1975). Total nacional y comunidades autónomas.
[25] Centro Nacional de Epidemiología, Incidencias acumuladas e indicadores de transmisibilidad (2021) COVID-19 (isciii.es).
[25] MoMo y CNE. Monitorización y Vigilancia diaria por todas las causas en España (2021). MoMo (isciii.es)

Glosario

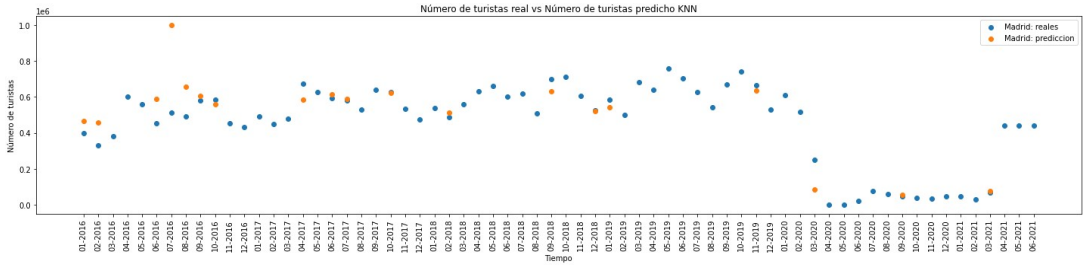
CCAA	Comunidad Autonoma
INE	Instituto nacional de estadistacas
OWID	Our World In Data
JHU	Johns Hopkins University
KNN	Algorimto k vecinos más cercanos
RF	Random Forest
MLP	Perceptrón multicapa
LR	Regresión lineal
Target	Objetivo

Anexos

A Gráficas de resultados adicionales aprendizaje automático

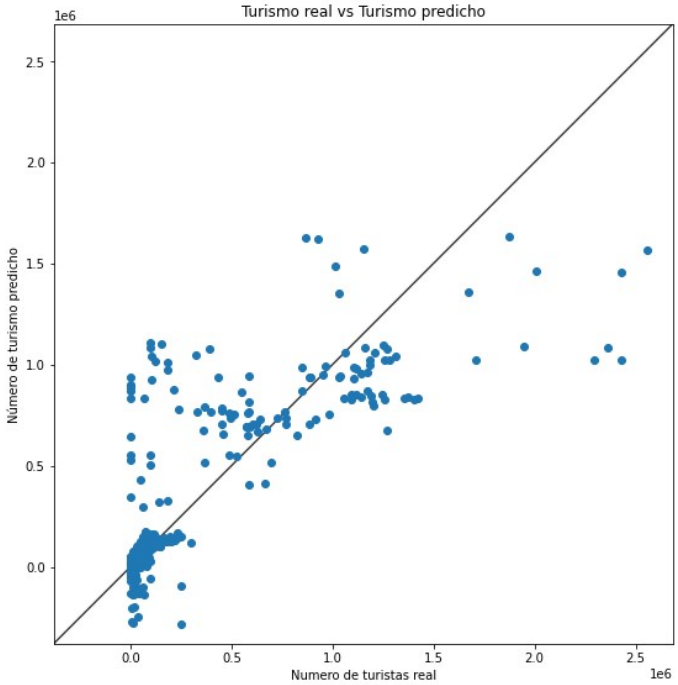
A.1 **Turismo España**

A.1.1 Resultados algoritmo KNN

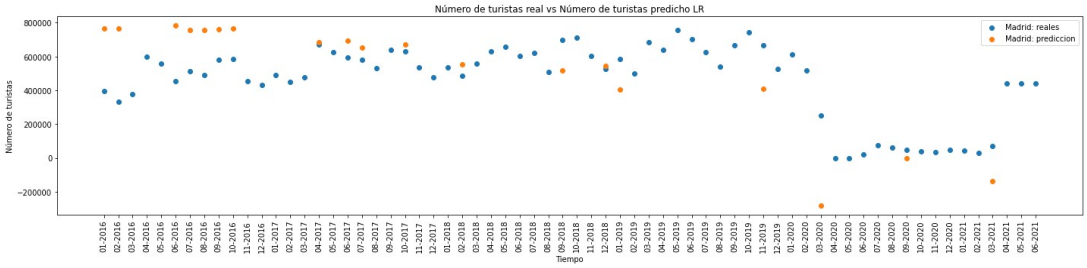


Gráfica A-0-1 Resultados prediciendo Madrid Abril 2021 KNN

A.1.2 Resultados Regresión Lineal

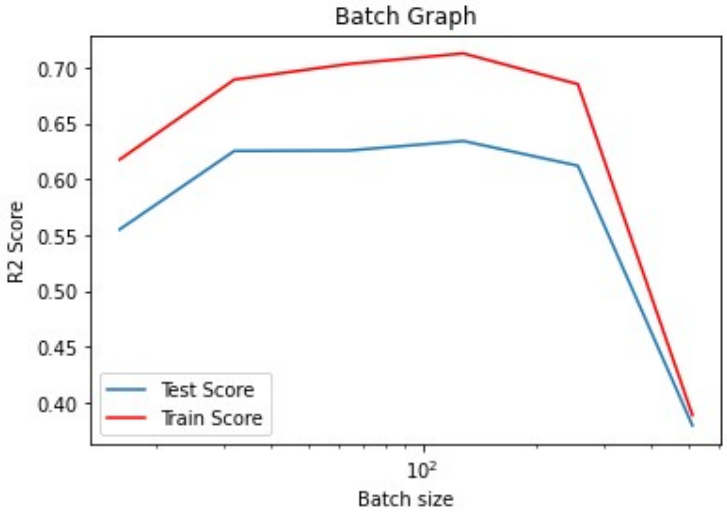


Gráfica A-0-2 Predicción turismo Test vs Predicho LR

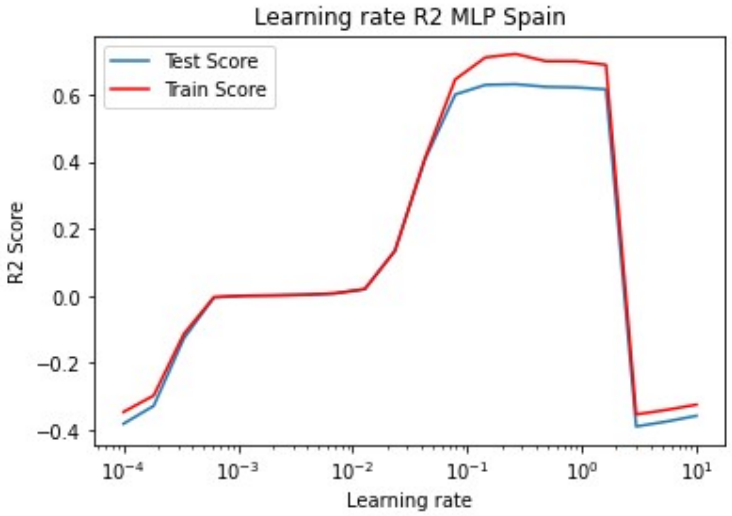


Gráfica A-0-3 Turistas en Madrid Real vs Predicho LR

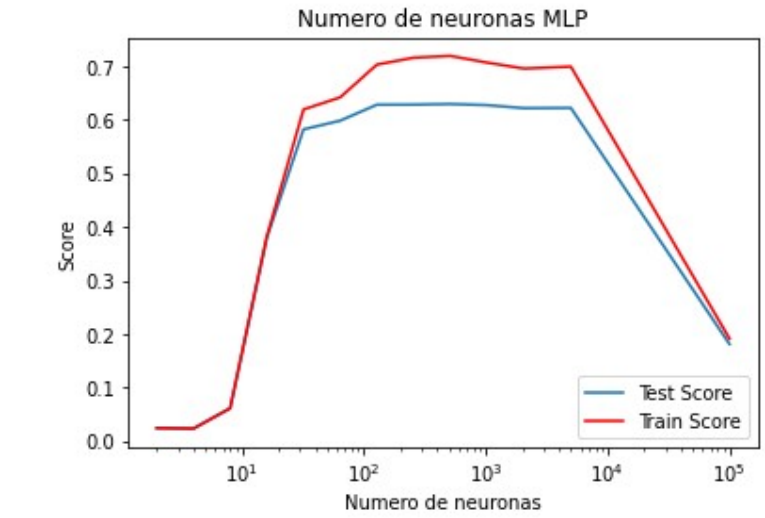
A.1.3 Resultados perceptrón multicapa



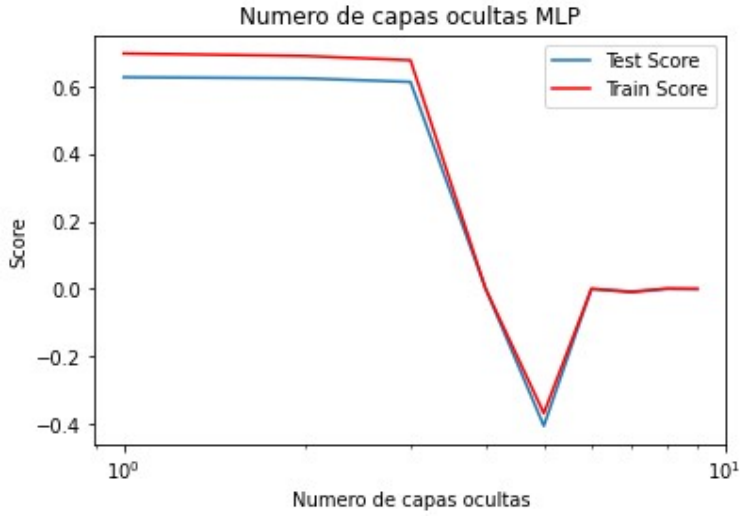
GráficaA- 0-4 MLP Turismo variando el tamaño de batch



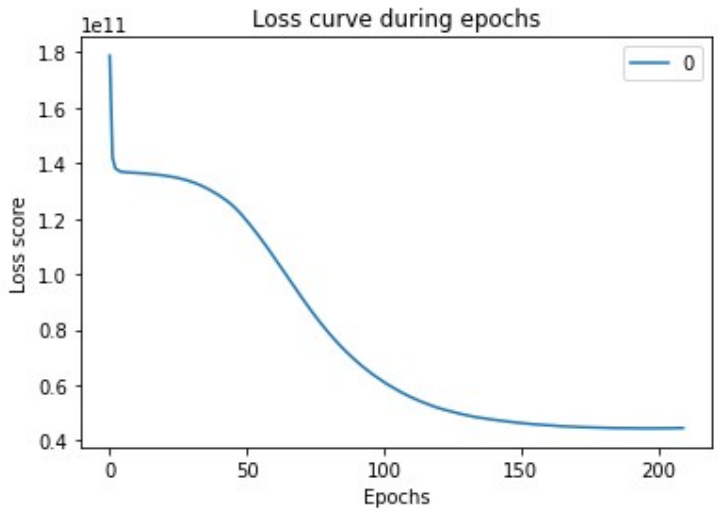
Gráfica A-0-5 MLP Turismo variando la tasa de aprendizaje



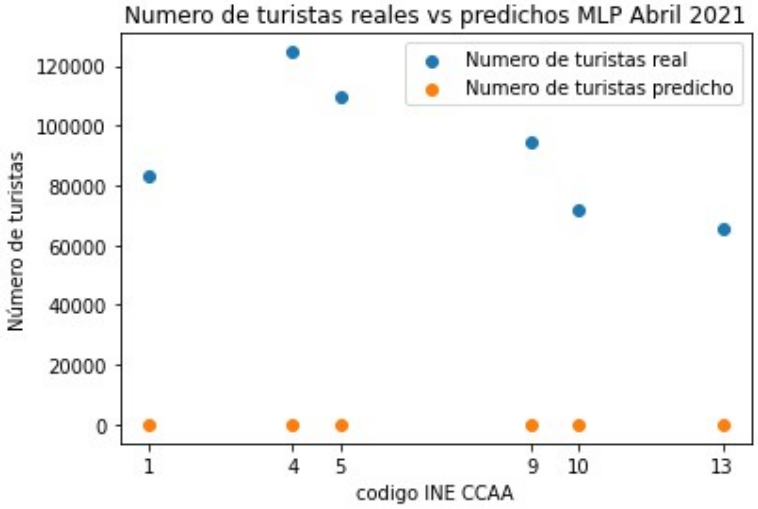
Gráfica A-0-6 MLP Turismo variando Número de neuronas



Gráfica A-0-7 MLP Turismo variando número de capas ocultas



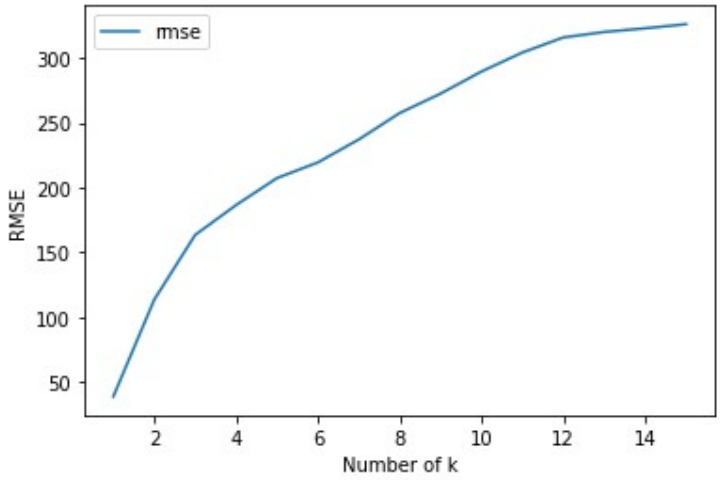
Gráfica A-0-8 MPL Turismo Curva de error aumentando épocas



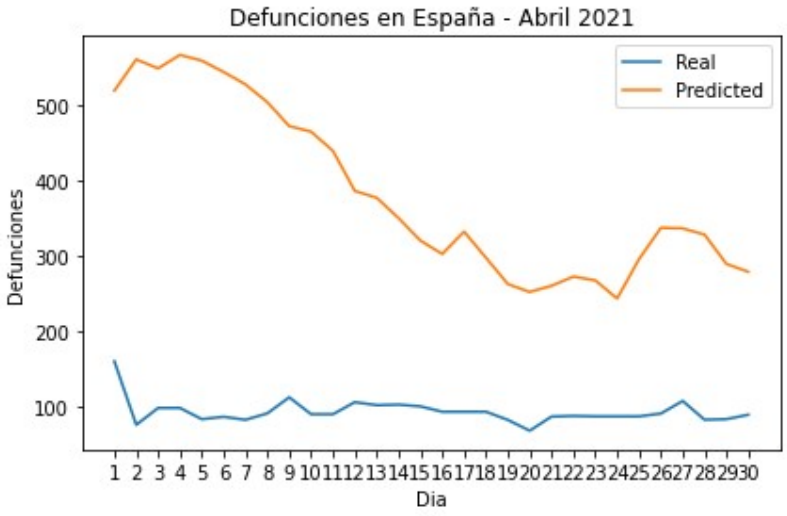
Gráfica A-0-9 MLP Turismo predicción vs real en 6 CCAA Abril 2021

A.2 **Defunciones Mundo**

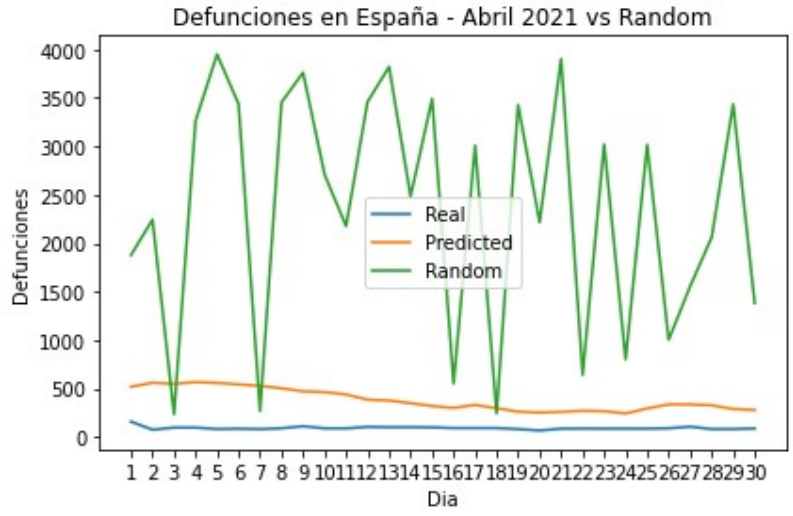
A.2.1 Resultados algoritmo KNN



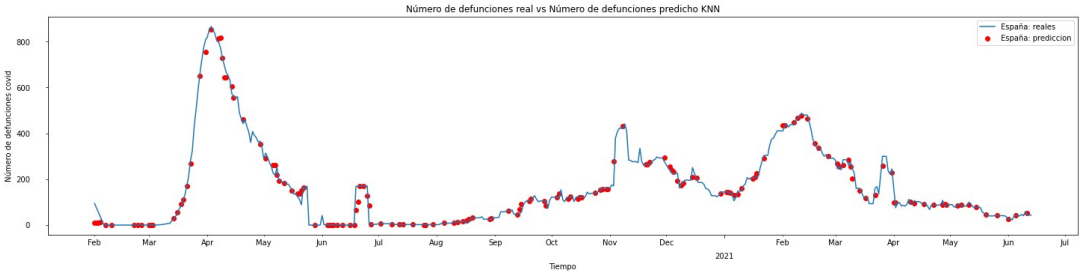
Gráfica A-0-10 RMSE variando k KNN-Mundo



Gráfica A-0-11 Defunciones España Abril KNN

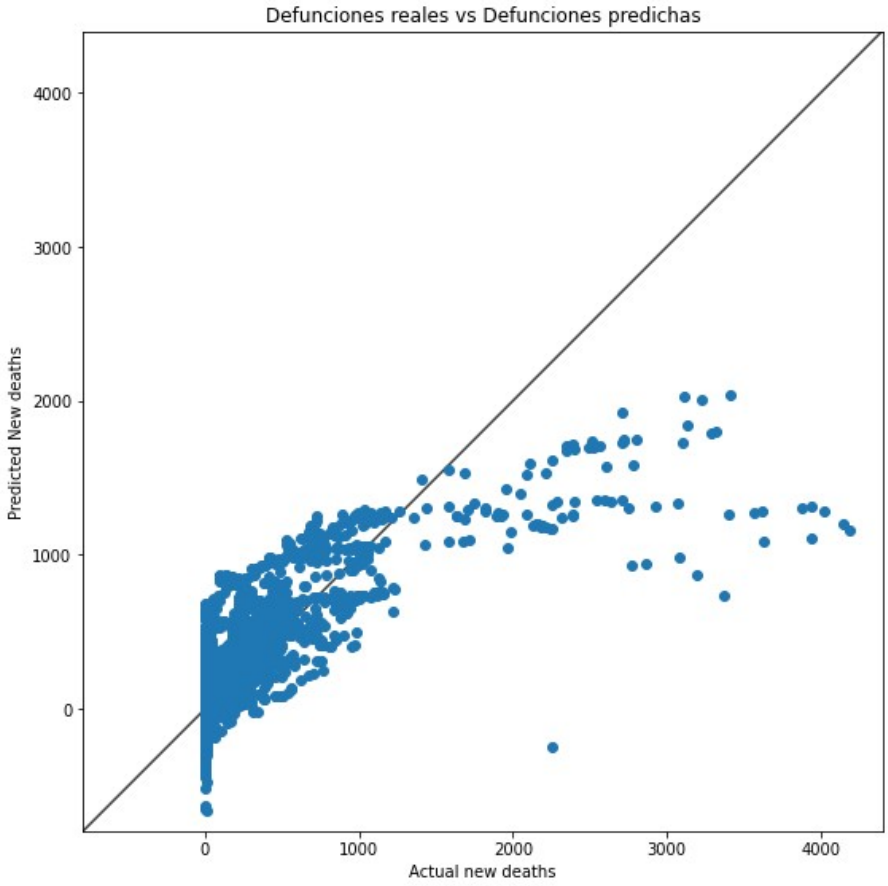


Gráfica A-0-12 Random vs KNN defunciones Abril 2021

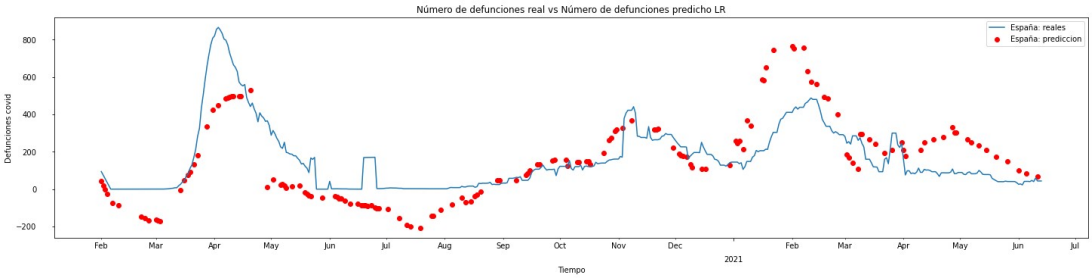


Gráfica A-0-13 Defunciones Covid en España KNN

A.2.2 Resultados COVID 19 Regresión lineal

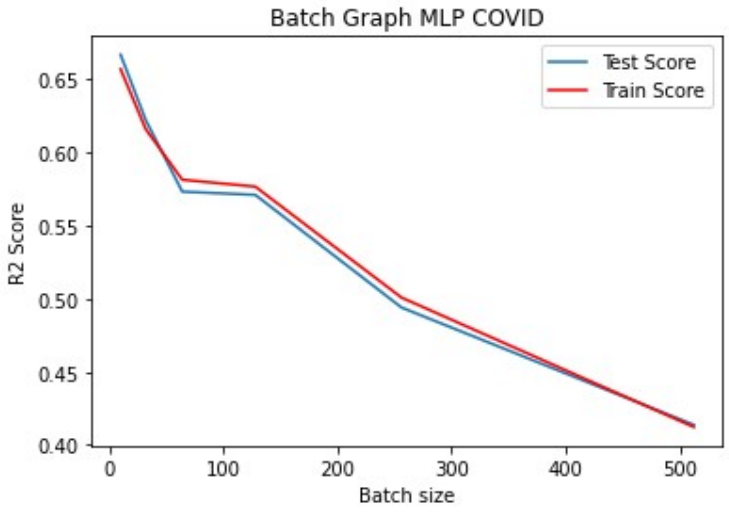


Gráfica A-0-14 Defunciones reales vs predichas LR COVID 19

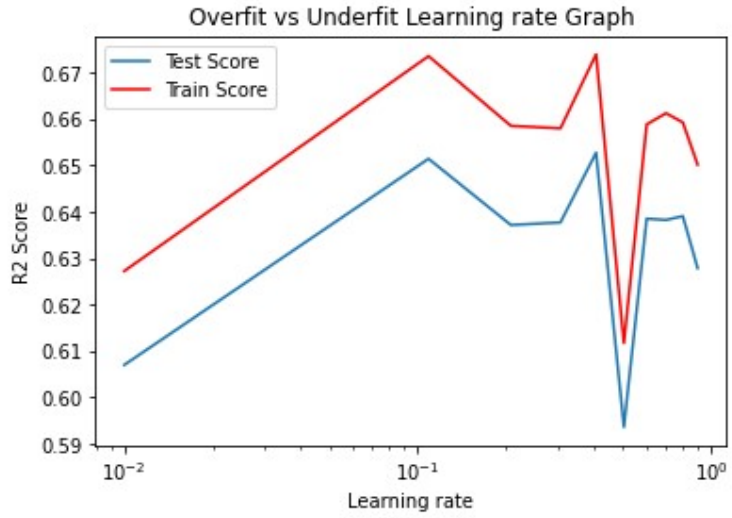


Gráfica A-0-15 Real vs Predichas defunciones España LR COVID-19

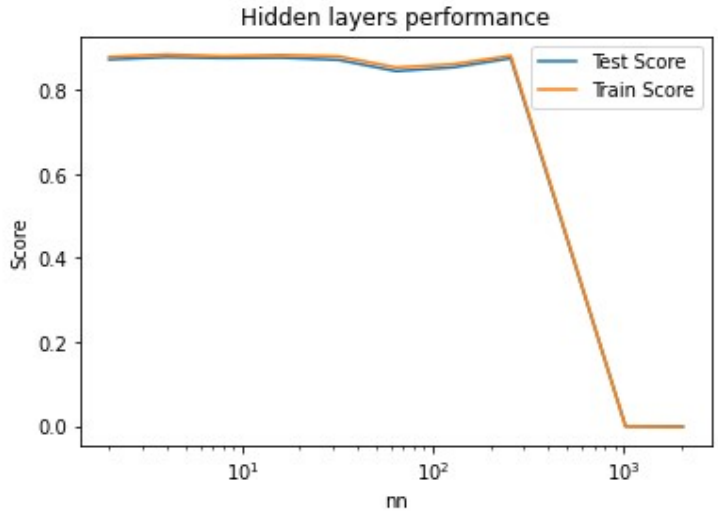
A.2.3 Resultados COVID 19 Perceptrón Multicapa



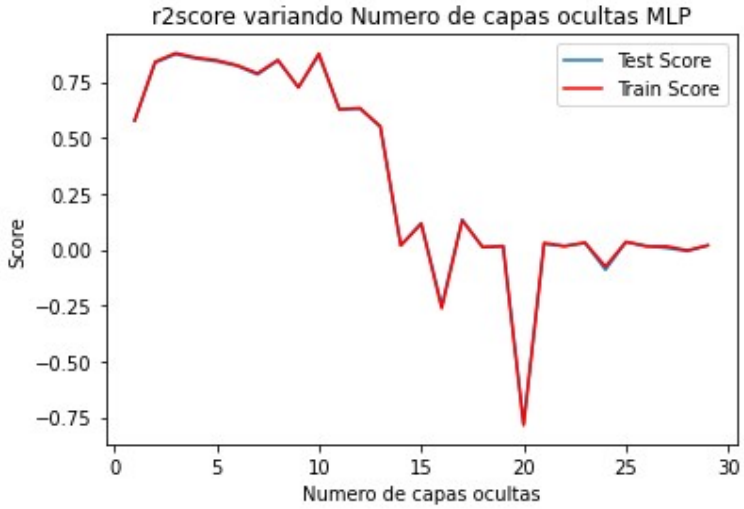
Gráfica A-0-16 MLP COVID-19 R2 score variando Batch Size



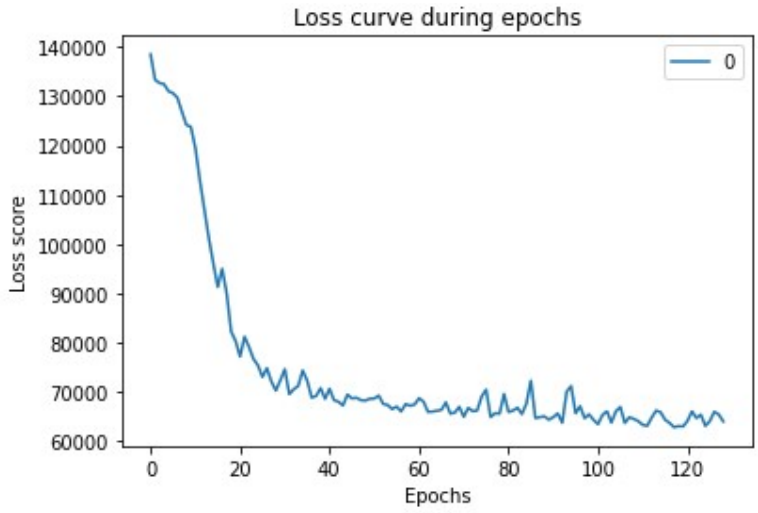
Gráfica A-0-17 MLP COVID 19 R2 score variando la tasa de aprendizaje



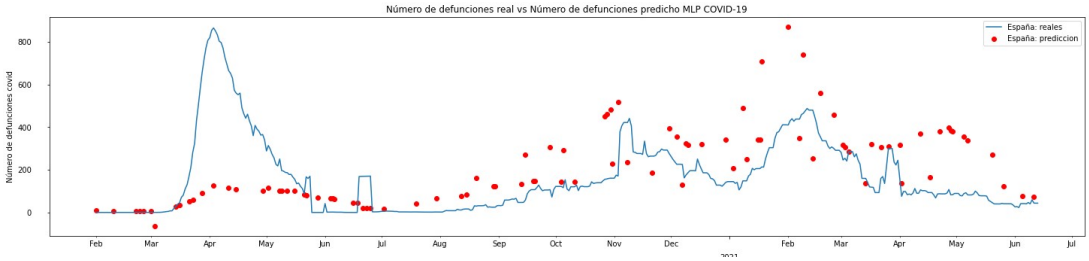
Gráfica A-0-18 MLP COVID-19 R2 score variando el número de neuronas en la capa oculta



Gráfica A-0-19 MLP COVID-19 R2 score variando Capas ocultas

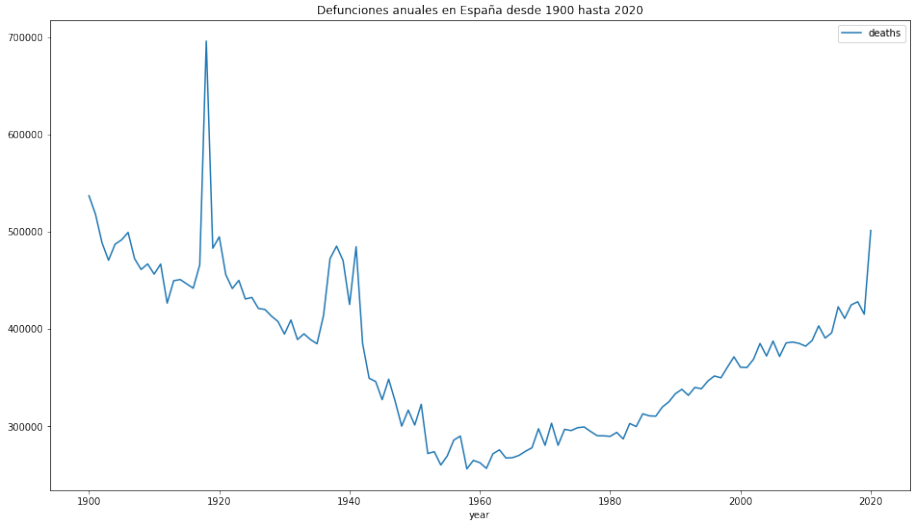


Gráfica A-0-20 MLP COVID-19 Loss curve aumentando epocas

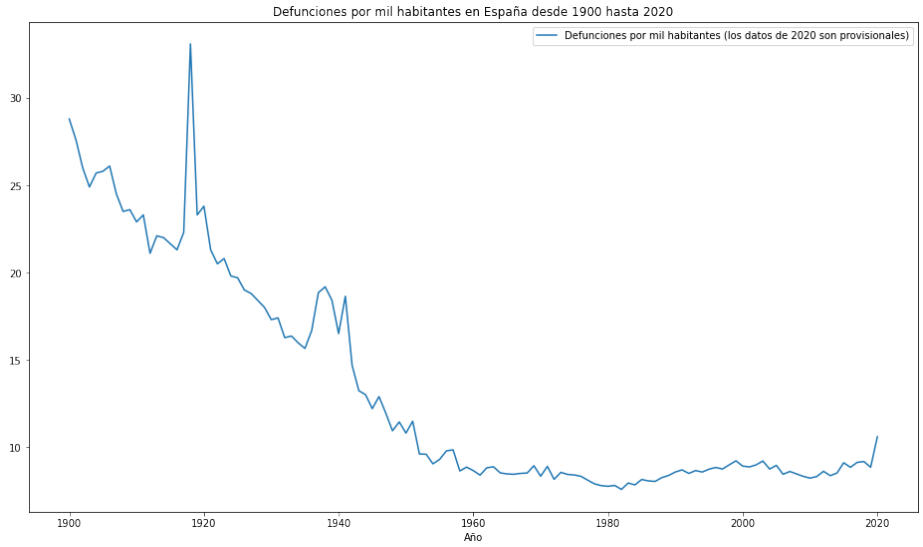


Gráfica A-0-21 MLP COVID-19 Real vs Predichas defunciones España

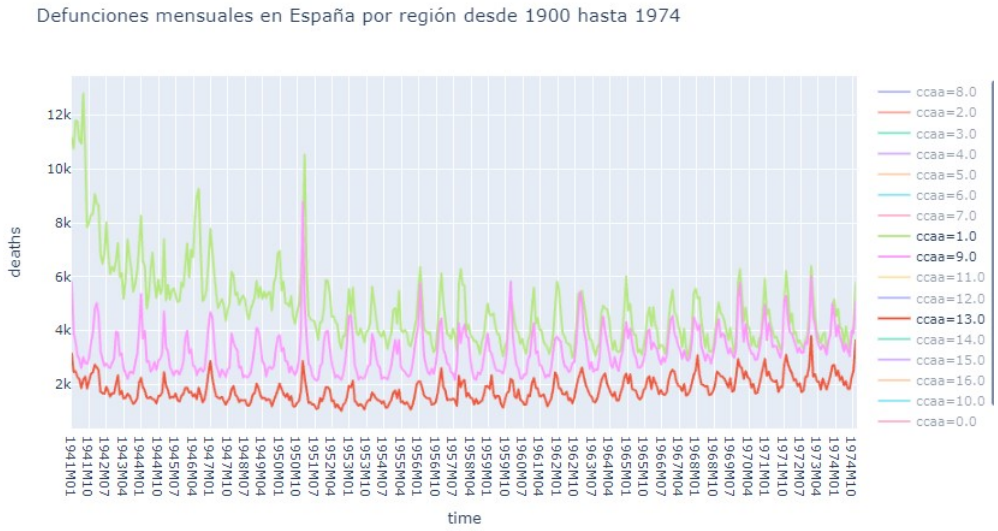
B Gráficas estudio previo



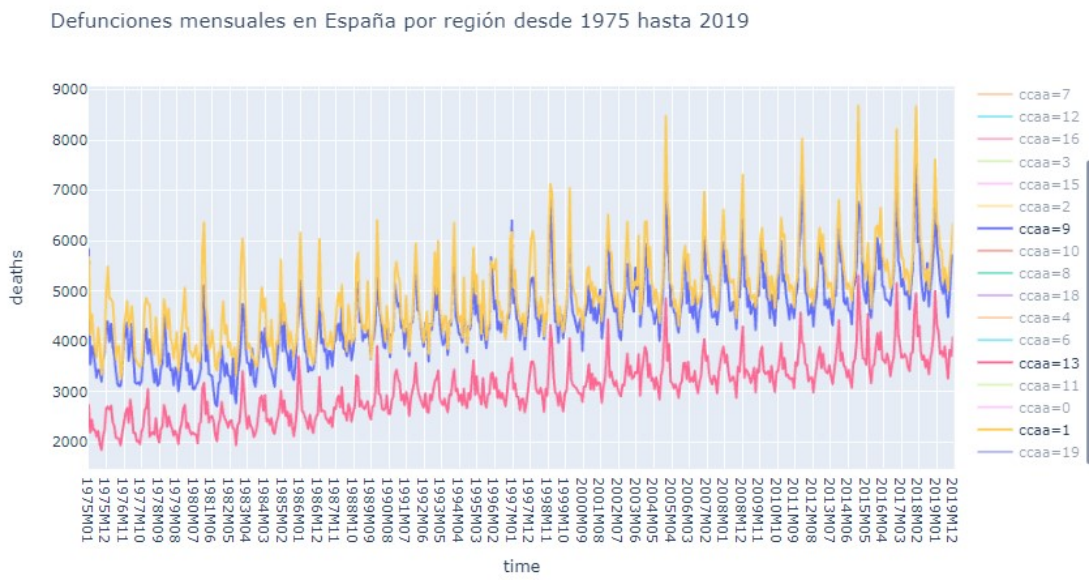
Gráfica B-0-22: Defunciones España desde 1900 hasta 2020



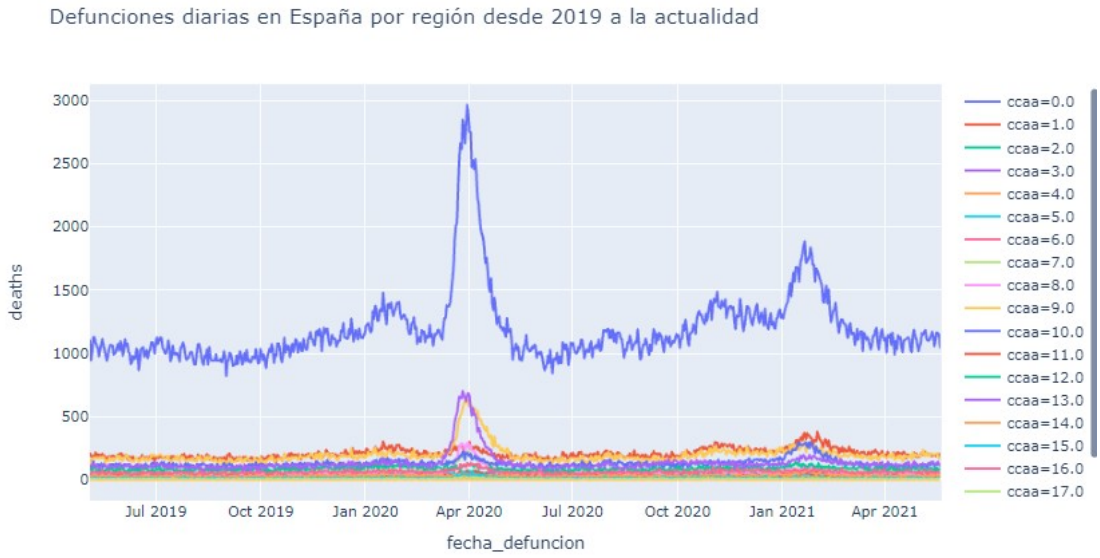
Gráfica B-0-23: Defunciones por mil habitantes desde 1900 hasta 2020



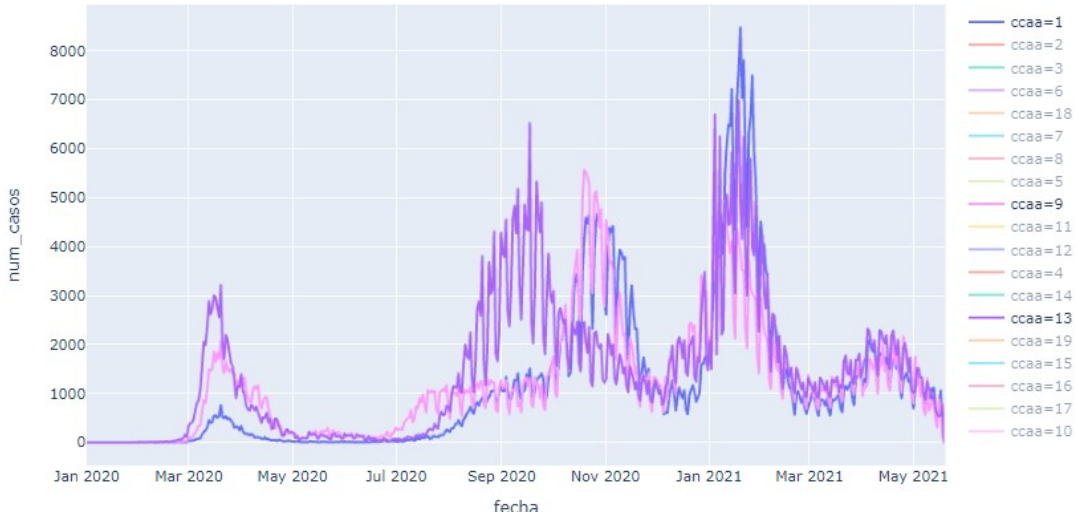
Gráfica B-0-24: Defunciones mensuales en Españas 1900-1974



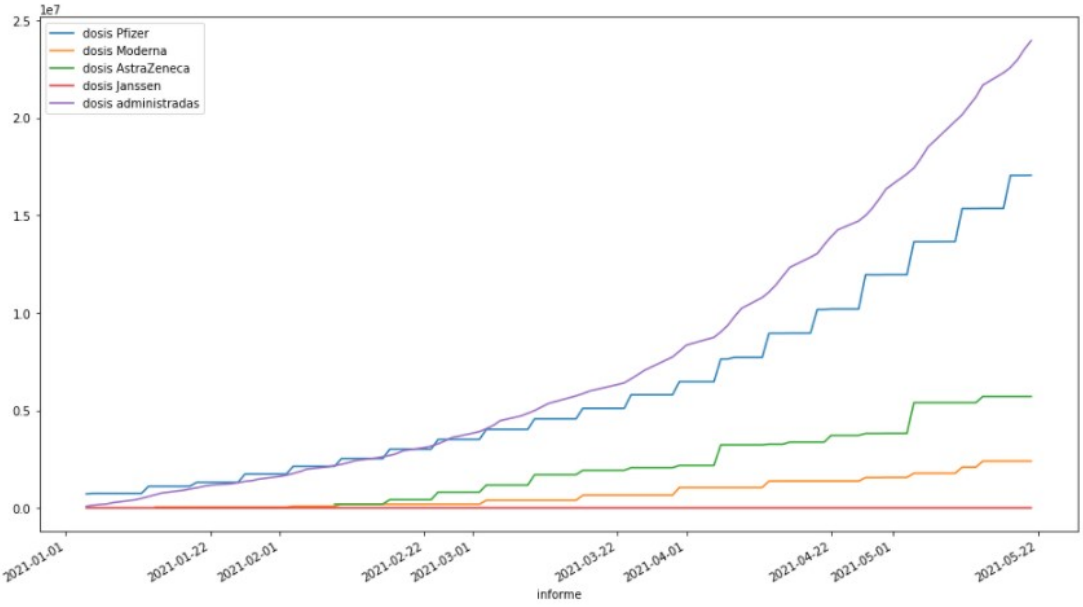
Gráfica B-0-25: Defunciones mensuales en España 1975-2019.



Gráfica B-0-26: Defunciones diarias desde 2019 a la actualidad.



Gráfica B-0-27: Casos COVID-19 en España



Gráfica B-0-28: Vacunación en España

C Códigos de identificación

	Literal
01	Andalucía
02	Aragón
03	Asturias, Principado de
04	Balears, Illes
05	Canarias
06	Cantabria
07	Castilla y León
08	Castilla - La Mancha
09	Cataluña
10	Comunitat Valenciana
11	Extremadura
12	Galicia
13	Madrid, Comunidad de
14	Murcia, Región de
15	Navarra, Comunidad Foral de
16	País Vasco
17	Rioja, La
18	Ceuta
19	Melilla